2019

Pattern discovery for genome-wide base composition evolution and genetic dissection of NDVI with UAV-based remote sensing in crops

Jinyu Wang
Iowa State University

Pattern discovery for genome-wide base composition evolution and genetic dissection of NDVI with UAV-based remote sensing in crops

by

Jinyu Wang

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Plant Biology

Program of Study Committee:
Jianming Yu, Major Professor
Yanhai Yin
Matthew Hufford
Jessica Barb
Peng Liu

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2019

TABLE OF CONTENTS

Page

NOMENCLATURE

CO	Crossover
CPDs	Cyclobutane Pyrimidine Dimers
DMRs	Differentially Methylated Regions
GWAS	Genome-Wide Association Studies
HTPPs	High-Throughput Phenotyping Platforms
NDVI	Normalized Difference Vegetation Index
PD	Private Domesticated SNPs
PI	Private Improved Cultivar SNPs
PL	Private Landrace SNPs
PW	Private Wild SNPs
TE	Transposable Element
UAV	Unmanned Aerial Vehicle
UV	Ultraviolet
WGBS	Whole-Genome Bisulfite Sequencing

ACKNOWLEDGMENTS

I would like to thank my major professor Dr. Jianming Yu for his support and guidance through my graduate study. He taught me not only skills on research and critical thinking, but also how to be confident and how to improve communication skills in the scientific community.

I also like to thank my Program of Study committee, Dr. Yanhai Yin, Dr. Matthew Hufford, Dr. Jessica Barb, and Dr. Peng Liu, for their valuable comments and suggestions on my research projects.

I want to thank Dr. Xianran Li in the group for his guidance during my graduate study. He gave me a lot of great suggestions for my research.

I also want to thank all other members in the Quantitative Genetics and Maize Breeding lab at Iowa State University, Dr. Tingting Guo, Dr. Adam Vanous, James McNellie, Qi Mu, Jialu Wei, Laura Tibbs, Gregory Schoenbaum, Paul White and former lab members Dr. Xiaoqing Yu, Dr. Xin Li, Dr. Matthew Dzievit, Yun Wu for their suggestions and help on my projects.

Finally, I would like to thank my family and friends for their support during the past six years.

ABSTRACT

Pattern discovery from biological data is crucial to advance our understanding of complex biological systems or biological processes and facilitate the application of our knowledge to benefit human needs. Base composition is an essential genomic feature. Findings of genome-wide base composition evolutionary pattern and its potential mechanisms can improve our understanding of genome evolution. Unmanned aerial vehicle-based high-throughput phenotyping platforms (UAV-HTPPs) can perform large-scale proximal measurements of phenotypic traits with high efficiency, high accuracy, and low cost, which provides novel opportunities to study the dynamic change of phenotypic traits across the growing season. The focus of my research is to study the genome-wide nucleotide evolutionary pattern following domestication in maize and soybean and time series normalized difference vegetation index (NDVI) data from a UAV-HTPP in maize.

We investigated the genome-wide base composition patterns through analyzing millions of SNPs segregating among 100 teosinte-maize accessions and among 302 wild-domesticated soybean accessions. Domesticated accessions have more nucleotide A and T across genome-wide polymorphic sites than wild accessions in maize and soybean. We demonstrated that different parts of the genome have differential contributions to the [AT]-increase between wild and domesticated accessions. The contribution to the [AT]-increase of non-genic part of the genome is greater than that of genic SNPs. The separation in [AT] values between wild and domesticated accessions is significantly enlarged in non-genic and pericentromeric regions. With motif frequency and sequence context analyses, we also showed that motifs (PyCG) related to solar-ultraviolet (UV) signature are enriched in non-genic and pericentromeric regions, particularly when they are methylated. Further genome scans using base-composition across polymorphic sites as a genome phenotype identify a set of putative candidate genes involved in

UV damage repair pathways. Our findings establish important connections among UV radiation, mutation, DNA repair, methylation, and genome evolution.

Time series NDVI from 5 critical growth stages of 1,752 diverse maize accessions were extracted from spectral images acquired with a UAV-HTPP. We analyzed the dynamic change of NDVI across the growing season. Genotypic differences were identified with clustering analysis of time series NDVI. We conducted genome-wide association studies (GWAS) using static NDVI values from individual time points and growth curve parameters of NDVI dynamics across the growing season. GWAS with both static NDVI values and growth curve parameters identified a number of association signals. Additionally, GWAS with model fitted NDVI values discovered the dynamic change of the SNP effect for trait-associated genetic loci, which likely suggests the role of gene-environment interplay in affecting the development of NDVI across the growing season. Our results indicate that UAV-based remote sensing can assist the genetic dissection of NDVI.

CHAPTER 1. GENERAL INTRODUCTION

Introduction

Plant domestication, the process in which wild plants were evolved into crop plants through artificial selection, plays a critical role in human history (Ross-Ibarra et al., 2007; Meyer et al., 2012). Substantial morphological changes occurred during domestication. A considerable number of studies have been conducted to understand the domestication process and the genes responsible for these morphological changes (Doebley et al., 2006; Purugganan and Fuller, 2009; Meyer and Purugganan, 2013; Olsen and Wendel, 2013). Genomes also went through profound changes during domestication. The advances in sequencing technology generated a huge amount of publicly available genomic data, which provides a great opportunity to study genome change as well as its potential mechanisms. Recent studies of DNA base-composition with populations separated by a domestication bottleneck event (Li et al., 2015b) and mutation rate with populations separated by a demographic bottleneck event (Harris, 2015) provided novel insights on genome evolution. DNA base composition, mutation spectrum, and the potential relationship between them need to be further investigated to advance our understanding of genome evolution.

DNA base composition is known to be associated with codon usage, phylogenic relatedness, and genome organization (Sueoka, 1962; Sharp and Matassi, 1994; Bernardi, 2000; Hershberg and Petrov, 2012; Costantini and Musto, 2017). Interestingly, a conserved base-composition pattern, modern accessions having more A and T nucleotides across genome-wide polymorphic sites than accessions sampled from corresponding progenitor species, was discovered across multiple species (Li et al., 2015b). However, little is known about the relative contribution of different genomic regions to this genome divergence pattern and its potential mechanisms.

The objectives of the first part of this dissertation are 1) to study the relative contribution of different genomic regions to the base-composition captured genome divergence pattern; 2) to study whether DNA polymorphisms occurred more frequently in certain sequence contexts; 3) to study whether there is mutation spectrum change during domestication; 4) to identify underlying genetic components of genome divergence in plant genomes.

Phenotyping under field conditions is critical for plant genetics, physiology, and agricultural research. However, phenotyping under field conditions is still time-consuming, labor-intensive, and error-prone, and it has been considered as a bottleneck for crop improvement (Cobb et al., 2013; Watanabe et al., 2017; Yang et al., 2017). Improved phenotyping efficiency and accuracy will greatly help researchers and breeders characterize the relationship between a plant's phenotype and genotype and assist the selection of high yielding variety. Equipped with high spatial and spectral resolutions of the sensors, unmanned aerial vehicle-based high-throughput phenotyping platforms (UAV-HTPPs) have high capacity and efficiency in conducting large-scale proximal filed measurements and crop condition monitoring, which provide a great solution to overcome the phenotyping bottleneck (Chapman et al., 2014; Liebisch et al., 2015; Haghighattalab et al., 2016).

Normalized difference vegetation index (NDVI) is the most popular product from UAV-based remote sensing. It is derived from the reflectance difference between the visible red spectral region (λ = 500-700 nm) and the near infrared region (NIR, (λ = 760–900 nm) (Kumar and Silva, 1973). Due to chlorophyll absorption, plants generally have low reflectance in the blue and red spectral portions. Therefore, NDVI is closely related to the leaf chlorophyll content and can successfully predict plant photosynthetic activity. NDVI is also known to be associated with many traits, including leaf area index (LAI), senescence, drought-adaptive traits, nitrogen usage efficiency, biomass, and grain yield (Duncan et al., 1967; Bort et al., 2005; Liebisch et al., 2015;

Condorelli et al., 2018). With a UAV-HTPP, NDVI measurements of a large diverse maize population at multiple growth stages across the growing season can be obtained, which provides a great opportunity to study the dynamic development of NDVI and perform genetic dissection of NDVI with genome-wide association study.

The objectives of the second part of this dissertation are 1) to study the dynamics of NDVI development with time series NDVI data obtained from a UAV-HTPP and 2) to statistically model the NDVI growth curve, and to conduct genetic dissection of NDVI using NDVI values observed from individual time points and NDVI growth curve parameters.

Dissertation Organization

This dissertation is organized into four chapters. Chapter 1 is the general introduction and literature review. Chapter 2 is the study of the evolutionary pattern of genome-wide base composition in maize and soybean. Chapter 3 is devoted to the study of time series NDVI data from a UAV-HTPP. These two chapters are written in the format of journal articles with their own abstract, introduction, materials and methods, results, discussion, and references. Chapter 4 is general conclusions.

Literature Review

DNA base composition

DNA consists of only four different nucleotides, namely A, T, C, and G. Except for some viruses, all of the living organisms have genomes consisting of exclusively DNA molecules (Koonin and Dolja, 2013). Therefore, DNA base composition, the percentage for each of the four nucleotides (A, T, C, and G) in a DNA sequence, is a fundamental genomic feature. A better understanding of the evolutionary pattern of DNA base composition can help us understand how genomes have changed over the evolution process.

Two important patterns about DNA base composition are discovered by Erwin Chargaff, which are described as Chargaff's first parity rule (PR1) and Chargaff's second parity rule

(PR2). PR1 holds that on a double-stranded DNA molecule, [A] = [T], and [C] = [G] (CHARGAFF et al., 1952). The validity of the PR1 constitutes the integral pre-requisite of Watson-Crick's double helix model. The less-known, PR2 holds that [A] ≈ [T], and [C] ≈ [G] on the individual strand of the double-stranded DNA molecule (Rudner et al., 1968; Mitchell and Bridge, 2006). Although PR2 was previously validated at the genome level (Mitchell and Bridge, 2006), until very recently, the validity of PR2 was demonstrated with individual-strand base composition across genome-wide single nucleotide polymorphisms (SNPs) using a population of related individuals (Li et al., 2015b). PR1 and PR2 are predominantly valid for genomes consisting of double-stranded DNA. And deviations may occur for single-stranded viral genomes (Albrecht-Buehler, 2006).

Base composition varies between genomes of different organisms but is more similar within closely related groups (Mooers and Holmes, 2000). Prokaryotic genomes have a base composition varying substantially with GC content ranging from 13-75% (Benson et al., 2015). Compared with prokaryotic genomes, eukaryotic genomes have a relatively narrow range of GC content, between 30-65% (Romiguier et al., 2010; Šmarda and Bureš, 2012; Benson et al., 2015). The varied GC content between genomes of different organisms can be used to draw phylogenetic inference (Mooers and Holmes, 2000). In plants, genomes of grass have generally higher GC content than that of other angiosperm families (King and Ingrouille, 1987; Barow and Meister, 2002).

Base composition can also be markedly variable within genomes. In general, coding regions are significantly more GC rich than non-coding regions (Aïssani and Bernardi, 1991; Bernardi, 1995; Glemin et al., 2014). Variations of GC content within genomes of complex organisms, such as plants and mammals, resulting in mosaic-like formation of large continuous DNA regions that are homogeneous in their GC content called isochores (Bernardi, 2000). In

both plants and mammals, it seems protein coding genes are concentrated in GC-rich isochores (Gardiner, 1996; Carels et al., 1998). Determination of GC content in specific regions of the genome can contribute to the discovery of gene-rich regions in the genome (Sumner et al., 1993). Isochores have not been identified in prokaryotic genomes that have relatively uniform GC content. Thus, isochores seem to represent a structural genome organization layer that is distinct to eukaryotes (Costantini and Musto, 2017). Indeed, it was shown that isochores are involved in chromosome packaging and high order genome structure (Jabbari and Bernardi, 2017).

Besides its impact on phylogenic relatedness and genome organization, base composition is also related to codon usage bias. Codon usage bias refers to the differences in frequency with which synonymous codons are used in coding DNA. Different organisms have different preferences for using a set of codons encode the same amino acid over the others (Athey et al., 2017). There are evidence showing that GC content is the driven factor for codon usage bias (Hershberg and Petrov, 2012). When synonymous codons are different in their proportion of G+C, GC content and codon usage are logically connected (Mooers and Holmes, 2000).

So far, nearly all of the previous findings of DNA base composition patterns were discovered through considering all the bases along DNA sequences within genomes. Until recently, a conserved base composition pattern was discovered with exclusively the dynamic part of the genome from populations of related individuals. Figure 1 shows that modern accessions have significantly higher [A] and [T] values across genome-wide polymorphic sites than accessions sampled from the corresponding progenitor species, and this base composition pattern was identified for multiple comparisons of accessions separated by domestication (Li et al., 2015b).

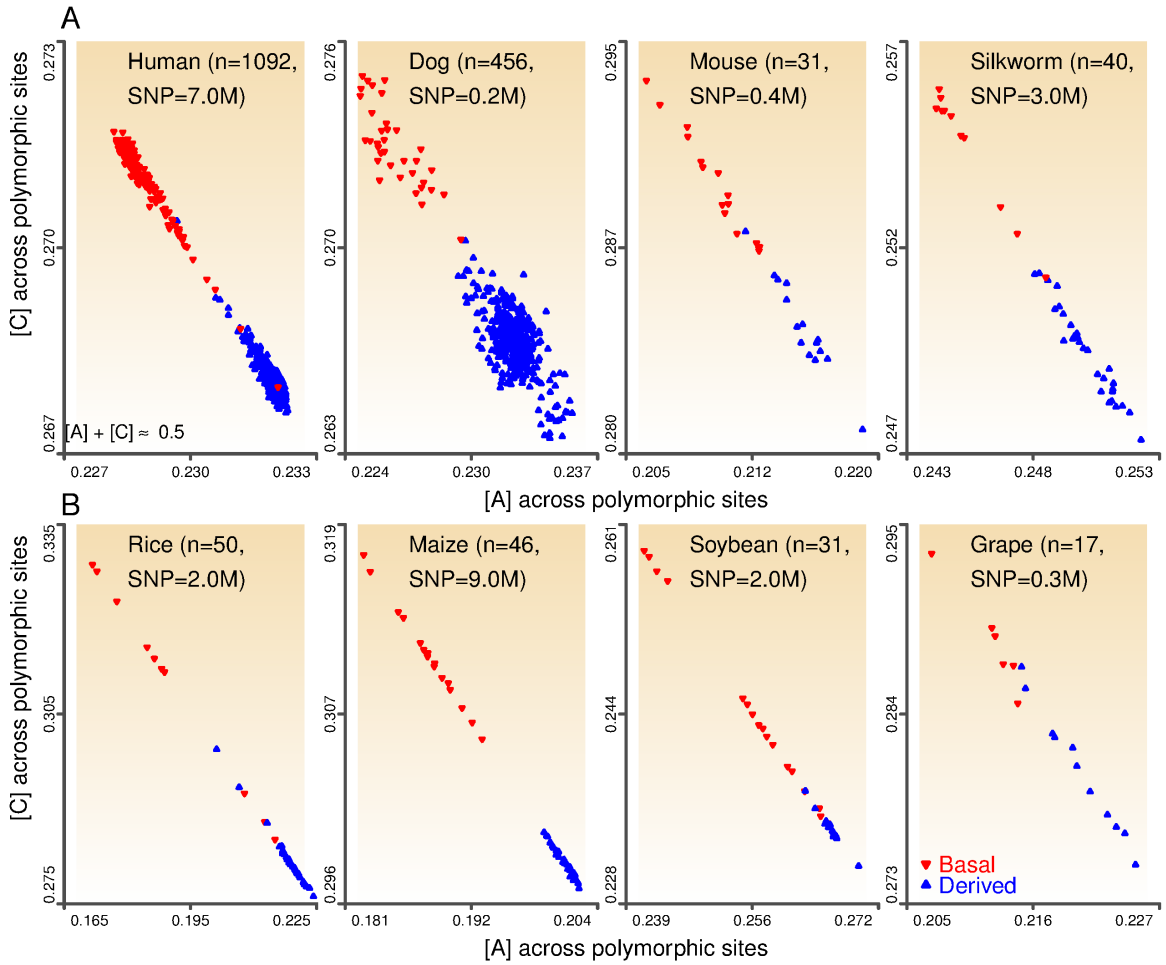


Figure 1. The base composition pattern summarized from genome-wide sequence polymorphisms in eight comparison sets (Li et al., 2015b).

Mutation

Mutation is a fundamental factor that generates the genetic variation upon which natural selection acts, and thus it plays a critical role in evolution. A mutation is the change of nucleotide sequence in the genome of an organism. Mutations can result from errors introduced during DNA replication, damage to DNA because of exposure to environmental factors such as solar light, radiation, and smoking, or insertion or deletion of DNA segments because of mobile genetic elements (Bertram, 2000; Aminetzach et al., 2005; Sharma et al., 2015). Mutations can alter DNA sequences in many different ways. Based on the effect on chromosome structure, mutations can be classified into large-scale mutations and small-scale mutations. Large-scale

mutations include duplications of chromosome segments, deletions of large chromosome regions, and chromosome rearrangement, and these mutations are likely to have more serious effects on organisms. Small-scale mutations include base substitutions, and small insertions and deletions, and most of these mutations have no or small effect on organisms. Base substitutions, referring to the exchange of a single nucleotide for another in a DNA sequence, are the most frequent type of mutations (Freese, 1959a). Base substitutions can be further classified into 4 transitions and 2 transversions. Transitions describe the exchange of a purine for a purine (A ↔ G) or a pyrimidine for a pyrimidine (C ↔ T), while transversions describe the exchange of a purine for a pyrimidine or a pyrimidine for a purine (A/G ↔ C/T) (Freese, 1959b).

Mutations occur according to certain biases. As the most frequent type of mutation, base substitution was shown to biased towards AT, and this is mostly due to the high rate of C/G to T/A transitions (Hershberg and Petrov, 2010; Lynch, 2010). The bacteria study in which mutational patterns were estimated with data from four diverse bacterial clades demonstrated that consistent across synonymous and non-synonymous sites mutations in all clades are consistently biased towards AT. Previous studies have illustrated that mutation rate vary across populations within the same species. A research in human reported the increased rate of TCC→TTC mutation in European population (Harris, 2015). Mutation also accumulated at different rates across populations. Accelerated rates of mutation accumulation were observed in non-Africans compared to Africans since divergence (Mallick et al., 2016). Indeed, divergence in the rate or type of mutations between populations are important factors affecting genetic variation patterns (Mathieson and Reich, 2017). Mutation rates also vary significantly among regions of the genome (Wolfe et al., 1989; Ellegren et al., 2003). It is shown that silent substitutions rate varies among genes and is associated with the base composition of genes as well as its flanking DNA.

Mutation bias is invoked as one of the hypotheses to explain base composition variation within and between species. Mutation bias was considered as the main force that determines the nucleotide content variation in bacteria (Muto and Osawa, 1987; Chen et al., 2004). A number of studies have shown that at least partial of the base composition can be attributed to mutation bias, especially because similar GC contents can be found in exon and intron regions of the same gene. Isochore structure is considered as the result of regional differences in mutational bias (Wolfe et al., 1989). A more important link was even found between base composition of the isochore and DNA replication process. While AT-rich segments seem to replicate late in the cell cycle, GC-rich segments seem to replicate early in the cell cycle (Tenzen et al., 1997). The relative effect of mutation rates and fixing probability on the pattern of human base composition was demonstrated with an evolutionary modeling research (Lipatov et al., 2006). A recent study showed that base composition difference can arise from mutation sites. By analyzing data from multiple spontaneous and induced mutation accumulation experiments, the study demonstrated that there are higher [AT] values across mutation sites in derived lines at the end of mutation experiments than in ancestral lines (Li et al., 2015b).

DNA repair

For most of the living organisms, DNA is the genetic information carrier which stores all the information for growth and development. Thus, it is vital to maintain genome integrity to assure the normal functionality of the organisms and to pass high fidelity sequence information to the next generation. Organisms have evolved elaborate mechanisms to respond to DNA damage caused by various endogenous and exogenous mutagens (Schärer, 2003; Bray and West, 2005). The responses to DNA damage include the activation of DNA repair pathways and activation of cell cycle checkpoints to inhibit cell proliferation transiently when the dosage of

DNA damage is low or transferring to programmed cell death when DNA damage is too severe to repair (Ciccia and Elledge, 2010; Hu et al., 2016a).

Depending on the type of DNA damage, different repair mechanisms will be applied to restore the lost information. When there are small lesions on DNA sequence, organisms use direct reversal or base excision repair (BER) to fix the damage. Small DNA lesions such as ultraviolet (UV) photolesions, alkylated bases, and methylation of guanine can be directly reversed in an error-free manner (Kato et al., 1994; Sancar, 2003; Yi and He, 2013). BER can correct forms of oxidative, deamination, and abasic single, non-helix-distorting base lesions to DNA (Dianov and Hübscher, 2013; Odell et al., 2013). If DNA lesions are bulky with multiple base damage, nucleotide excision repair (NER), mismatch repair (MMR), or translesion synthesis (TLS) is used to address the DNA damage. NER will be the choice to remove bulky helix-distorting DNA lesions such as cyclobutane pyrimidine dimers (CPDs) from UV radiation (Ikehata and Ono, 2011). MMR is a post replicative repair pathway that is typically used to fix base mismatches occurred during replication and contribute significantly to replication fidelity (Kunkel, 2009). TLS is a DNA damage tolerate process which can replicate past DNA lesions with TLS polymerase in a relatively low fidelity manner (Fuchs and Fujii, 2013). When more severe DNA damage like strand-breaks happen, mechanisms like single stranded break repair (SSBR) for repairing single-strand breaks (SSBs) and non-homologous end joining (NHEJ) and homologous-recombination (HR) for repairing double-strand breaks (DSBs) will be activated (Caldecott, 2008; Panier and Boulton, 2014). SSBR is used to fix SSBs that are mostly generated from oxidative damage, abasic sites, or incorrect activity of the DNA topoisomerase 1 (Wang, 2002; Caldecott, 2008). DSBs are highly toxic to organisms as they can cause genome rearrangements. The NHEJ requires short homologous sequences at the single-stranded tails of the DNA ends to be joined and may further introduce mutations during the repair process(Panier

and Boulton, 2014). The HR relies on homologous sequences for template-directed DNA repair synthesis in a high-fidelity manner (Li and Heyer, 2008).

An effective DNA repair system is critical to maintain genome integrity and assure the organism's normal function and development. Whereas, some nucleotide changes that can escape the DNA monitoring and repairing systems are necessary to provide genetic variations for the evolutionary process. A recent study in human showed that DNA repair genes are enriched surrounding loci associated with the base-composition variation, which may suggest the potential role of DNA repair systems in the long-term genome evolution (Li et al., 2015b). The divergence of DNA repair genes between populations separated by population bottleneck merit further study to advance our understanding of its role in genome evolution.

Plant genome is exposed to various exogenous mutagens, such as solar-UV radiation, reactive oxygen species, excess boron or aluminum, and pathogenic microorganisms (Hu et al., 2016b). Solar-UV radiation is one of the major exogenous mutagens to plants as plants use solar light for photosynthesis and UV is a component of solar light. In general, UV-radiation can be classified into 3 classes based on the wavelength: UVA (315-400 nm), UVB (280-315 nm), and UVC (100-280 nm). Because UVC is blocked by the ozone layer and atmosphere, the UV component of solar light reaching the earth's surface consists of only UVA and UVB. Solar-UV radiation can induce various DNA lesions, such as cyclobutane pyrimidine dimers (CPDs) and 6-4 pyrimidine-pyrimidone photoproducts (Ikehata and Ono, 2011). It is known that solar UV induces CPDs, the primary solar UV-induced DNA lesion, preferentially at 5-methylcytosine-containing dipyrimidine sites (5′-Py_mCG-3′) and results in C→T base transitions, which is termed as solar-UV signature (Ikehata and Ono, 2007). CPDs distort the DNA's double-helix structure and consequently influence DNA unwinding and DNA replication, which ultimately affect cell cycle (Nawkar et al., 2013). Plant cells encode ataxia telangiectasia mutated (ATM)

and ATM- and RAD3-related (ATR) pathways that can sense DNA modifications and active cell cycle arrest in response to solar UV-induced damage (Hu et al., 2016a). CPDs can be by-passed through TLS with DNA polymerases like Polη, Rev1, be repaired by photolyase with energy from blue light, or be excised and repaired through NER (Landry et al., 1997; Liu et al., 2000; Takahashi et al., 2005; Anderson et al., 2008).

DNA methylation in plants

DNA methylation, the process by which a methyl group is added to the cytosine base of DNA to form 5-methylcytosine, is a dominating form of epigenetic modification that is crucial to gene regulation and genome stability (Robertson, 2005; Slotkin and Martienssen, 2007). Functions of DNA methylation include regulating gene expression, silencing transposons and repeat sequences, gene imprinting, and chromosome interactions (Zhang et al., 2018). In plants, DNA methylation occurs in all cytosine sequence contexts: CG, CHG and CHH (where H = A, C, or T). DNA methylation is conserved in plants. While DNA methylation is maintained by DNA methyltransferase, active DNA demethylation in plants requires DNA demethylase and involves the direct removal of the 5-mC base with 5-mC DNA glycosylases and a BER pathway. De novo DNA methylation in plants is mediated by RNA directed DNA methylation pathway (RdDM) (Matzke and Mosher, 2014).

DNA methylation is involved in many biological processes. Disruption of DNA methylation can result in developmental abnormalities in plants. For example, disrupted DNA methylation can inhibit tomato fruit ripening (Lang et al., 2017). One of the major functions of DNA methylation is regulating gene expression. DNA methylation at the promoter regions of genes mostly inhibits gene transcription through either inhibiting the binding of transcription activators or promoting the binding of transcription repressors (Zhang et al., 2006). DNA methylation at gene body regions could inhibit aberrant transcription from internal cryptic

promoters (Takuno and Gaut, 2012). Another major function of DNA methylation is transposon silencing. Transposon activity can affect genome stability as the relocation of transposons and insertion of transposable elements cause mutations to the genome. Pericentromeric heterochromatin and some of the transposon/repeat-containing regions in *A. thaliana* are heavily methylated (Zhang et al., 2018). RdDM maintained asymmetric (CHH) methylation is critical for transposon silencing (Li et al., 2015a). DNA hypomethylation will facilitate transposon mobilization (La et al., 2011). DNA methylation is involved in chromosome interaction through influencing the epigenetic state of chromatin (Feng et al., 2014). DNA methylation may prevent potential chromosome interactions at the KNOT structure, and it is also a major epigenetic determinant of chromosome interactions in pericentromeric regions in plants. In addition, DNA methylation also plays important roles in plant growth and development as well as the response to abiotic and biotic stress (Secco et al., 2015; Hewezi et al., 2017; Lang et al., 2017).

DNA methylation also relates to the varied mutation rate along chromosomes. Methylation of cytosine at the CpG dipyrimidine sites is prone to C to T spontaneous deamination (Ehrlich and Wang, 1981). More interestingly, it is shown that solar UV-induced CPD formation was significantly enhanced as a result of the methylation of cytosine at CpG sites, which indicates the role of DNA methylation in solar-UV induced mutagenesis (Tommasi et al., 1997). Many studies have shown that the relative frequency of DNA methylation in all contexts varies substantially along the chromosome (Song et al., 2013; West et al., 2014; Springer and Schmitz, 2017). DNA methylation is primarily distributed in pericentromeric heterochromatin regions that are mostly composed of tandem repeats and transposons. Considering the varied DNA methylation level along plant chromosomes and the potential effect of DNA methylation on solar-UV induced mutagenesis, it will be interesting to see if solar-UV induced mutations vary along chromosomes.

Phenotyping bottleneck and phenomics

The world population is projected to rise to 9.7 billion by 2050. In order to meet the future demand of food and fiber from the increasing world population, crop yields need to increase at an annual rate of 2.4%, but the current growth rate is only 1.3% (Fischer and Edmeades, 2010). Given climate change, recurring drought events, and limited agricultural land, the challenge is exacerbated by the necessity to accelerate research to develop high-yield and stress-tolerant crop varieties. Genetic improvement of crop plays a key role in improving crop production. However, the rates of genetic improvement of many crops are still lagging behind what is required to meet the future demand (Ray et al., 2013).

The major challenge for genetic improvement of crops is connecting genotypes with phenotypes so that the yield potential of a genotype can be realized in a given environment (White et al., 2012). It is possible to obtain high-yield and stress-tolerant crops to improve agricultural production if we can better understand the connection between genotypes and phenotypes. Over the past two decades, there are has been significant progress in molecular profiling and sequencing technologies, which leads to the development of many genomic resources and molecular technologies in crop plants. With these resources and technologies, researchers and breeders had found numerous applications such as marker-assisted breeding and genomic selection to increase the breeding efficiency in crops (Cobb et al., 2013).

Compared to the genotyping technologies, phenotyping technologies are not improved at the competitive pace to facilitate the connection between genotypes and phenotypes. The critical value of phenotyping has long been recognized in the genetic improvement of crops. Best genotypes were select based on their phenotypes for a very long time. Precise phenotyping has been addressed with experimental design, adequate size of multi-environment trials, and uniform management and cultivation practices. But traditional phenotyping methods are invasive, labor-

intensive and time-consuming, and mostly deal with one or a few traits at a specific time point, which makes it difficult to perform thorough functional analysis to connect genotypes and phenotypes. It becomes more challenging with the requirement of sampling multiple environments trials and a large sample size to conduct genetic dissection of complex quantitative traits. Phenotyping has become the bottleneck for crop improvement (Chapman et al., 2014).

Phenomics is the acquisition of appropriate multi-dimensional phenotypic data at multiple levels of organization, aiming to have a more complete characterization of phenotypic space instructed by a particular genome or set of genomes (Houle et al., 2010; Dhondt et al., 2013). Phenomics is considered as a natural complement to genomics to advance biology from a few aspects: a) phenomics enables us to trace causal relationships between genotypes and phenotypes; b) phenomics assists the genetic dissection of complex traits; c) phenomics allows us to give causal explanations at the phenotypic level. Plant phenotype is the result of genotype, environment, and their interactions, and it can change from time to time and from environment to environment (Houle et al., 2010).

Plant phenomics research integrates knowledge from many different disciplines, including agronomy, life science, mathematics, engineering, and computer science to explore the multi-dimensional phenotype information of plant growth. Driven by technological advancements in imaging sensors, robotics, and software pipelines, impressive progress has been made in plant phenomics. For example, a smartphone platform was developed for field phenotyping through image taking and image analysis at organ level (Confalonieri et al., 2017); the X-ray micro-computed topography (CT) was introduced into the 3D imaging of maize roots (Pan et al., 2017); High resolution 3D scanners were used to obtain morphological structure of plant organs (Rist et al., 2018).

High-throughput phenotyping platforms (HTPPs)

High-throughput phenotyping (HTP) is an assessment of plant phenotypes on a large scale and high speed, which is not achievable with traditional phenotyping methods (Dhondt et al., 2013). High-throughput phenotyping platforms (HTPPs) are important tools for crop phenomics.

In recent years, many HTPPs have been developed, and some of them are automated facilities in greenhouse or growth chambers with precise environmental control (Yang et al., 2017). These HTPPs are automated and high-precise, which greatly improve data collection efficiency and accuracy. However, most of these HTPPs have high construction and maintenance costs, which makes it inaccessible to many research institutions and limits its applications (Kolukisaoglu and Thurow, 2010). Besides, although the HTPPs in controlled environments enable researchers to capture detailed, non-destructive information throughout the plant growth cycle, it's hard to translate the genetic information identified within controlled environments into phenotype information under the field conditions. Field conditions are extremely heterogeneous and complex. Thus, the results from controlled environments can be very different from the actual situations that plants will experience in the field (Araus and Cairns, 2014).

Field-based phenotyping is critical for genetic improvement of crops as it measures the ultimate result of genetic factors, environmental factors, and the interaction between them (Araus and Cairns, 2014). Over the last few years, there has been increased interest in field-based phenotyping platforms (FBPPs) that use rigid motorized gantry, ground wheeled vehicles, or aerial vehicles, with a wide range of cameras and sensors, to acquire comprehensive phenotypic data in field conditions (Furbank and Tester, 2011; Fritsche-Neto and Borém, 2015; Walter et al., 2015). Phenobot, an auto-steered and self-propelled FBPPs equipped with RGB cameras was developed for measuring plant architecture parameters in biomass sorghum (Fernandez et al.,

2017). FIELD SCANAYZERS, a rigid motorized gantry based FBPPs mounted with multiple types of cameras, sensors, and illuminating systems were developed for high-throughput monitoring of crop performance (Virlet et al., 2017). There is no doubt that these ground-based phenotyping platforms have significantly improved our phenotyping ability, but they have low efficiency in measuring a large number of plots at different field locations, and some of them are unable to measure different crop systems (Haghighattalab et al., 2016).

Aerial-based phenotyping platforms allow the rapid characterization of many plots and monitoring of large-scale crop performance, overcoming one of the major limitations of ground-based phenotyping platforms (Chapman et al., 2014). In recent years, remarkable progress has been made for unmanned aerial vehicle-based HTPPs (UAV-HTPPs), which are powerful remote sensor-bearing platforms for various agricultural applications (Haghighattalab et al., 2016; Kyratzis et al., 2017; Condorelli et al., 2018; Han et al., 2018). For example, UAV-HTPPs were used for vegetation indices, plant height, and canopy cover measurement in maize (Han et al., 2018). UAV-HTPPs provide a low-cost approach to meet the critical requirements of high spatial, spectral, and temporal resolutions. UAV-HTPPs are able to cover the entire experiment in a very short time, performing rapid characterization for a large number of plots while minimizing the effect of varied environmental conditions.

UAV-HTPPs mainly consist of two major components: an unmanned aerial vehicle and a sensor. The typical unmanned aerial vehicles (UAVs) used for UAV-HTPPs are multi-rotors, helicopters, fixed-wing, blimps, and fly wing (Espinoza et al., 2015). The most frequently used UAVs for FBP are multi-rotor UAVs as they have the advantage of low cost, low requirements for taking off and landing, and hover ability. Whereas, multi-rotor UAVs have some limitations, such as relatively short flight time, lower payload, and sensitivity to weather (Peña et al., 2013). Different type of sensors, such as RGB camera, multispectral cameras, hyperspectral sensors,

infrared thermal imagers, and light detection and ranging (LIDAR) are being used on UAV-HTPPs for different remote sensing purposes (Liebisch et al., 2015; Ludovisi et al., 2017; Madec et al., 2017; Condorelli et al., 2018). One type of sensors commonly used by UAV are multispectral cameras. Multispectral cameras are capable of sensing and recording both invisible and visible parts of the electromagnetic spectrum, which can be used to obtain the spectral absorption and reflectance characteristics of crops; they have been widely deployed to evaluate both biological and physiological characteristics of a crop, to monitor crop growth, and to predict crop yield (Øvergaard et al., 2010; Candiago et al., 2015; Kyratzis et al., 2017).

Remote sensing is a non-destructive and resource conservative technique, which can obtain information about an object without making physical contact with it. With the advantage for data collection, satellite or aerial remote sensing have been widely used precision agriculture, such as monitoring soil properties (Ge et al., 2011), assessing biotic and abiotic stress (Gao, 1996; Mirik et al., 2011), and estimating yield or biomass levels (Serrano et al., 2000; Shanahan et al., 2001). These remote sensing methods provide spatial information for a large area, but they are unsuitable for obtaining high resolution images required for plant phenotyping. Remote sensing with low altitude UAV is able to acquire high spatiotemporal resolution images that promptly provide precise field conditions (Shi et al., 2016). A recent study that directly compares three remote sensing methods including UAV, satellite-based imagery, proximal sensing demonstrated that UAV based remote sensing performed best for measuring canopy temperature and NDVI in plant breeding (Tattaris et al., 2016). Remotely sensed data can be collected at varying scales and resolutions to complement in situ data, including adaptation to water stress, vegetation indices, leaf area index (LAI), chlorophyll measurements, and yield potential (Weber et al., 2012; Yang et al., 2017).

NDVI

One type of remotely sensed data is different vegetation indices. There has been a lot of progress in the development and application of remotely sensed vegetation indices since the last half century. The rationale behind the various applications of these vegetation indices is that combinations of different spectral-bands are able to reveal information such as photosynthetic capacity, vegetation cover, leaf water content, and nitrogen deficiency (Jensen, 2007).

The improved understanding of the plants' spectral properties facilitated the application of vegetation indices (Moss and Loomis, 1952; Gates et al., 1965; Kumar and Silva, 1973). Plant leaf structure determines how vegetation interacts with sunlight. And the amount of photosynthetic pigments, mostly chlorophyll, contained in a leaf affects its total amount of absorbed solar radiation. Green leaves have strong absorption in the blue and red portion (λ = 500-700 nm) of the spectrum, and less so in the green portion of the spectrum. That is why leaves appear to be green to our eyes. Sunlight in the near infrared (NIR) portion (λ = 760–900 nm) of the spectrum is strongly reflected from the surface of leaf due because of the cellular structure and the air cell wall-protoplasm-chloroplast interfaces (Kumar and Silva, 1973).

NDVI is the most widely used vegetation index. It is calculated from the reflectance measurements in the red and NIR portion of the spectrum:

$$NDVI = (R_{NIR} - R_{red})/(R_{NIR} + R_{red})$$

where R_{NIR} and R_{red} are the reflectances in visible red and NIR portion of the spectrum respectively. While green leaves reflect less visible light and more NIR, yellow or aging leaves reflect a larger portion of visible light and less NIR. Since NDVI combines the reflectance characteristics of both red and NIR portion of the spectrum, it is able to predict photosynthetic capacity (Govaerts and Verhulst, 2010).

NDVI is known to be associated with chlorophyll content and leaf area index (LAI) (Broge and Leblanc, 2001; Gitelson et al., 2003). NDVI has been related to biomass, grain yield, nutrient deficiency, drought-adaptive traits, and stay-green and senescence (Duncan et al., 1967; Bort et al., 2005; Liebisch et al., 2015; Condorelli et al., 2018). Many studies have shown that NDVI is associated with drought-adaptive traits, and many quantitative trait loci (QTLs) underlying NDVI were identified through genetic dissection of NDVI as a proxy for drought adaptive traits in wheat (Bowman et al., 2015; Condorelli et al., 2018). NDVI at milk-stage was found to be strongly positively correlated with final yield and biomass in durum wheat (Marti et al., 2007). In maize, NDVI at flowering was most correlated to final grain yield, and this period showed the best potential for predicting grain yield. (Robert et al., 1999; Spitkó et al., 2016). NDVI is very closely related to the N content of leaves (Raun et al., 2001), and it has been used to study the nitrogen deficiency or nitrogen usage efficiency in crops (Cabrera-Bosquet et al., 2011; Vergara-Díaz et al., 2016). NDVI is also related to stay-green and senescence (Liebisch et al., 2015; Duan et al., 2017). Since senescence is a dynamic process, genotype exhibiting different senescence development trends could have a similar final NDVI (Christopher et al., 2014).

NDVI values change during the plant growing season. Plants exhibit different features as they develop across the growing season, and the differences in many of these features can be captured by NDVI. Time series NDVI data obtained from satellite remote sensing imagery have been used to develop NDVI response curves (Also referred as crop phenological curves) across the growing season to monitor or estimate crop growth under various climate conditions (Masialeti et al., 2010; Wang et al., 2016). Different crops have different NDVI response curves due to their differences in the timing of green-up, peak greenness, and senescence. But in general annual crops such as maize, soybean, wheat, and sorghum have bell-shape like NDVI response

curves: NDVI values of these crops first increase at a slow rate after plants' emergence, increase at a fast rate during rapid vegetative growth stage, reach peak values close to or after vegetation completion, and decrease because of senescence (Masialeti et al., 2010).

Maize is one of the most widely cultivated cereals worldwide (FAOSTAT, 2019). Further improvement of maize production through advanced breeding methods and comprehensive phenotyping with the high-throughput phenotyping platforms is critical to meet the future demand of food. In recent years, NDVI data obtained from the UAV-based phenotyping platforms have been applied in maize for monitoring water stress, managing nitrogen usage, and estimating vigor and yield (Vergara-Díaz et al., 2016; Nasir and Tharani, 2017; Wahab et al., 2018). No study in maize has yet reported the use of NDVI values obtained from UAV-based remote sensing for genetic dissection of NDVI. It will be very interesting to conduct genome-wide association studies for NDVI data obtained from UAV-based remote sensing platforms in a large diverse maize population.

References

Aïssani B, Bernardi G (1991) CpG islands, genes and isochores in the genomes of vertebrates. Gene 106: 185-195

Albrecht-Buehler G (2006) Asymptotically increasing compliance of genomes with Chargaff's second parity rules through inversions and inverted transpositions. Proc Natl Acad Sci U S A 103: 17828-17833

Aminetzach YT, Macpherson JM, Petrov DA (2005) Pesticide resistance via transposition-mediated adaptive gene truncation in Drosophila. Science 309: 764-767

Anderson HJ, Vonarx EJ, Pastushok L, Nakagawa M, Katafuchi A, Gruz P, Di Rubbo A, Grice DM, Osmond MJ, Sakamoto AN, et al (2008) Arabidopsis thaliana Y-family DNA polymerase eta catalyses translesion synthesis and interacts functionally with PCNA2. Plant J 55: 895–908

Araus JL, Cairns JE (2014) Field high-throughput phenotyping: the new crop breeding frontier. Trends Plant Sci 19: 52–61

Athey J, Alexaki A, Osipova E, Rostovtsev A, Santana-Quintero L V., Katneni U, Simonyan V, Kimchi-Sarfaty C (2017) A new and updated resource for codon usage tables. BMC Bioinformatics 18: 391

Barow M, Meister A (2002) Lack of correlation between AT frequency and genome size in higher plants and the effect of nonrandomness of base sequences on dye binding. Cytometry 47: 1-7

Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2015) GenBank. Nucleic Acids Res 43: D30-D35

Bernardi G (1995) The human genome: organization and evolutionary history. Annu Rev Genet 29: 445-476

Bernardi G (2000) Isochores and the evolutionary genomics of vertebrates. Gene 241: 3–17

Bertram JS (2000) The molecular biology of cancer. Mol Aspects Med 21: 167-223

Bort J, Casadesus J, Nachit MM, Araus JL (2005) Factors affecting the grain yield predicting attributes of spectral reflectance indices in durum wheat: growing conditions, genotype variability and date of measurement. Int J Remote Sens 26: 2337-2358

Bowman BC, Chen J, Zhang J, Wheeler J, Wang Y, Zhao W, Nayak S, Heslot N, Bockelman H, Bonman JM (2015) Evaluating grain yield in spring wheat with canopy spectral reflectance. Crop Sci 55: 1881-1890

Bray CM, West CE (2005) DNA repair mechanisms in plants: crucial sensors and effectors for the maintenance of genome integrity. New Phytol 168: 511-528

Broge NH, Leblanc E (2001) Comparing prediction power and stability of broadband and hyperspectral vegetation indices for estimation of green leaf area index and canopy chlorophyll density. Remote Sens Environ 76: 156-172

Cabrera-Bosquet L, Molero G, Stellacci A, Bort J, Nogués S, Araus J (2011) NDVI as a potential tool for predicting biomass, plant nitrogen content and growth in wheat genotypes subjected to different water and nitrogen conditions. Cereal Res Commun 39: 147-159

Caldecott KW (2008) Single-strand break repair and genetic disease. Nat Rev Genet 9: 619-631

Candiago S, Remondino F, De Giglio M, Dubbini M, Gattelli M (2015) Evaluating multispectral images and vegetation indices for precision farming applications from UAV images. Remote Sens 7: 4026–4047

Carels N, Hatey P, Jabbari K, Bernardi G (1998) Compositional properties of homologous coding sequences from plants. J Mol Evol 46: 45-53

Chapman SC, Merz T, Chan A, Jackway P, Hrabar S, Dreccer MF, Holland E, Zheng B, Ling TJ, Jimenez-Berni J (2014) Pheno-copter: a low-altitude, autonomous remote-sensing robotic helicopter for high-throughput field-based phenotyping. Agronomy 4: 279–301

Chargaff E, Lipshitz R, Green C (1952) Composition of the desoxypentose nucleic acids of four genera of sea-urchin. J Biol Chem 195: 155-160

Chen SL, Lee W, Hottes AK, Shapiro L, McAdams HH (2004) Codon usage between genomes is constrained genome-wide mutational processes. Proc Natl Acad Sci U S A 101: 3480-3485

Christopher JT, Veyradier M, Borrell AK, Harvey G, Fletcher S, Chenu K (2014) Phenotyping novel stay-green traits to capture genetic variation in senescence dynamics. Funct Plant Biol 41: 1035-1048

Ciccia A, Elledge SJ (2010) The DNA damage response: making it safe to play with knives. Mol Cell 40: 179-204

Cobb JN, DeClerck G, Greenberg A, Clark R, McCouch S (2013) Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype-phenotype relationships and its relevance to crop improvement. Theor Appl Genet 126: 867–887

Condorelli GE, Maccaferri M, Newcomb M, Andrade-Sanchez P, White JW, French AN, Sciara G, Ward R, Tuberosa R (2018) Comparative aerial and ground based high throughput phenotyping for the genetic dissection of NDVI as a proxy for drought adaptive traits in durum wheat. Front Plant Sci 9: 893

Confalonieri R, Paleari L, Foi M, Movedi E, Vesely FM, Thoelke W, Agape C, Borlini G, Ferri I, Massara F, et al (2017) PocketPlant3D: analysing canopy structure using a smartphone. Biosyst Eng 164: 1-12

Costantini M, Musto H (2017) The isochores as a fundamental level of genome structure and organization: A general overview. J Mol Evol 84: 93-103

Dhondt S, Wuyts N, Inze D (2013) Cell to whole-plant phenotyping: the best is yet to come. Trends Plant Sci 18: 433–444

Dianov GL, Hübscher U (2013) Mammalian base excision repair: the forgotten archangel. Nucleic Acids Res 41: 3483-3490

Doebley JF, Gaut BS, Smith BD (2006) The molecular genetics of crop domestication. Cell 127: 1309–1321

Duan T, Chapman SC, Guo Y, Zheng B (2017) Dynamic monitoring of NDVI in wheat agronomy and breeding trials using an unmanned aerial vehicle. F Crop Res 210: 71-80

Duncan WG, Williams WA, Loomis RS (1967) Tassels and the productivity of maize 1. Crop Sci 7: 37-39

Ehrlich M, Wang RYH (1981) 5-Methylcytosine in eukaryotic DNA. Science 212: 1350-1357

Ellegren H, Smith NGC, Webster MT (2003) Mutation rate variation in the mammalian genome. Curr Opin Genet Dev 13: 562-568

Espinoza CZ, Sankaran S, Pumphrey MO, Miklas PN, Carter AH, Vandemark GJ, Knowles NR, Khot LR, Jarolmasjed S, Sathuvalli VR (2015) Low-altitude, high-resolution aerial imaging systems for row and field crop phenotyping: a review. EUR J AGRON 70: 112-123

FAOSTAT (2019) FAOSTAT: Statistical database. http://faostat.fao.org/

Feng S, Cokus SJ, Schubert V, Zhai J, Pellegrini M, Jacobsen SE (2014) Genome-wide Hi-C analyses in wild-type and mutants reveal high-resolution chromatin interactions in Arabidopsis. Mol Cell 55: 694-707

Fernandez MGS, Bao Y, Tang L, Schnable PS (2017) A high-throughput, field-based phenotyping technology for tall biomass crops. Plant Physiol 174: 2008–2022

Filipski J (1987) Correlation between molecular clock ticking, codon usage, fidelity of DNA repair, chromosome banding and chromatin compactness in germline cells. FEBS Lett 217: 184-186

Fischer RA, Edmeades GO (2010) Breeding and cereal yield progress. Crop Sci 50: S85-S98

Freese E (1959a) The difference between spontaneous and base-analogue induced mutations of Phage T4. Proc Natl Acad Sci U S A 45: 622-633

Freese E (1959b) The specific mutagenic effect of base analogues on Phage T4. J Mol Biol 1: 87-105

Fritsche-Neto R, Borém A (2015) Phenomics: How Next-Generation Phenotyping is Revolutionizing Plant Breeding. Springer, New York City, USA

Fuchs RP, Fujii S (2013) Translesion DNA synthesis and mutagenesis in prokaryotes. Cold Spring Harb Perspect Biol 5: a012682

Furbank RT, Tester M (2011) Phenomics - technologies to relieve the phenotyping bottleneck. Trends Plant Sci 16: 635–644

Gao BC (1996) NDWI - A normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sens Environ 58: 257-266

Gardiner K (1996) Base composition and gene distribution: critical patterns in mammalian genome organization. Trends Genet 12: 519-524

Gates DM, Keegan HJ, Schleter JC, Weidner VR (1965) Spectral properties of plants. Appl Opt 4: 11-20

Ge Y, Thomasson JA, Sui R (2011) Remote sensing of soil properties in precision agriculture: A review. Front Earth Sci 5: 229-238

Gitelson AA, Viña A, Arkebauer TJ, Rundquist DC, Keydan G, Leavitt B (2003) Remote estimation of leaf area index and green leaf biomass in maize canopies. Geophys Res Lett 30: 1248

Glemin S, Clement Y, David J, Ressayre A (2014) GC content evolution in coding regions of angiosperm genomes: a unifying hypothesis. Trends Genet 30: 263–270

Govaerts B, Verhulst N (2010) The normalized difference vegetation index (NDVI) Greenseeker(TM) handheld sensor: toward integrated evaluation of crop management part B-user guide. CIMMYT, Mexico

Haghighattalab A, Perez LG, Mondal S, Singh D, Schinstock D, Rutkoski J, Ortiz-Monasterio I, Singh RP, Goodin D, Poland J (2016) Application of unmanned aerial systems for high throughput phenotyping of large wheat breeding nurseries. Plant Methods 12:35

Han L, Yang G, Yang H, Xu B, Li Z, Yang X (2018) Clustering field-based maize phenotyping of plant-height growth and canopy spectral dynamics using a UAV remote-sensing approach. Front Plant Sci 9:1638

Harris K (2015) Evidence for recent, population-specific evolution of the human mutation rate. Proc Natl Acad Sci U S A 112: 3439–3444

He XJ, Chen T, Zhu JK (2011) Regulation and function of DNA methylation in plants and animals. Cell Res 21: 442-465

Hershberg R, Petrov DA (2010) Evidence that mutation is universally biased towards AT in bacteria. PLoS Genet 6: e1001115

Hershberg R, Petrov DA (2012) On the limitations of using ribosomal genes as references for the study of codon usage: a rebuttal. PLoS One 7: e49060

Hewezi T, Lane T, Piya S, Rambani A, Rice JH, Staton M (2017) Cyst nematode parasitism induces dynamic changes in the root epigenome. Plant Physiol 174: 405-420

Houle D, Govindaraju DR, Omholt S (2010) Phenomics: the next challenge. Nat Rev Genet 11: 855–866

Hu Z, Cools T, De Veylder L (2016a) Mechanisms used by plants to cope with DNA damage. Annu Rev Plant Biol 67: 439–462

Ikehata H, Ono T (2011) The mechanisms of UV mutagenesis. J Radiat Res 52: 115–125

Ikehata H, Ono T (2007) Significance of CpG methylation for solar UV-induced mutagenesis and carcinogenesis in skin. Photochem Photobiol 83: 196–204

Jabbari K, Bernardi G (2017) An isochore framework underlies chromatin architecture. PLoS One 12: e0168023

Jensen JR (2000) Remote Sensing of Environment: An Earth Resource. Saddle River, NJ, USA

Kato T, Todo T, Ayaki H, Ishizaki K, Morita T, Mitra S, Ikenaga M (1994) Cloning of a marsupial DNA photolyase gene and the lack of related nucleotide sequences in placental mammals. Nucleic Acids Res 22: 4119-4124

King GJ, Ingrouille MJ (1987) DNA base composition heterogeneity in the grass genus Briza L. . Genome 29: 621-626

Kolukisaoglu Ü, Thurow K (2010) Future and frontiers of automated screening in plant sciences. Plant Sci 178: 476-484

Koonin E V., Dolja V V. (2013) A virocentric perspective on the evolution of life. Curr Opin Virol 3: 546-557

Kumar R, Silva L (1973) Light ray tracing through a leaf cross section. Appl Opt 12: 2950-2954

Kunkel TA (2009) Evolving views of DNA replication (in)fidelity. Cold Spring Harb Symp Quant Biol 74: 91-101

Kyratzis AC, Skarlatos DP, Menexes GC, Vamvakousis VF, Katsiotis A (2017) Assessment of vegetation indices derived by UAV imagery for durum wheat phenotyping under a water limited and heat stressed Mediterranean environment. Front Plant Sci 8: 1114

La H, Ding B, Mishra GP, Zhou B, Yang H, Bellizzi MDR, Chen S, Meyers BC, Peng Z, Zhu JK, et al (2011) A 5-methylcytosine DNA glycosylase/lyase demethylates the retrotransposon Tos17 and promotes its transposition in rice. Proc Natl Acad Sci U S A 108: 15498-15503

Landry LG, Stapleton AE, Lim J, Hoffman P, Hays JB, Walbot V, Last RL (1997) An Arabidopsis photolyase mutant is hypersensitive to ultraviolet-B radiation. Proc Natl Acad Sci U S A 94: 328–332

Lang Z, Wang Y, Tang K, Tang D, Datsenka T, Cheng J, Zhang Y, Handa AK, Zhu JK (2017) Critical roles of DNA demethylation in the activation of ripening-induced genes and inhibition of ripening-repressed genes in tomato fruit. Proc Natl Acad Sci U S A 114: E4511-E4519

Li Q, Gent JI, Zynda G, Song JW, Makarevitch I, Hirsch CD, Hirsch CN, Dawe RK, Madzima TF, McGinnis KM, et al (2015a) RNA-directed DNA methylation enforces boundaries between heterochromatin and euchromatin in the maize genome. Proc Natl Acad Sci U S A 112: 14728–14733

Li X, Heyer WD (2008) Homologous recombination in DNA repair and DNA damage tolerance. Cell Res 18: 99-113

Li X, Scanlon MJ, Yu J (2015b) Evolutionary patterns of DNA base composition and correlation to polymorphisms in DNA repair systems. Nucleic Acids Res 43: 3614–3625

Liebisch F, Kirchgessner N, Schneider D, Walter A, Hund A (2015) Remote, aerial phenotyping of maize traits with a mobile multi-sensor approach. Plant Methods 11: 9

Lin JY, Jacobus BH, SanMiguel P, Walling JG, Yuan Y, Shoemaker RC, Young ND, Jackson SA (2005) Pericentromeric regions of soybean (Glycine max L. Merr.) chromosomes consist of retroelements and tandemly repeated DNA and are structurally and evolutionarily labile. Genetics 170: 1221–1230

Lipatov M, Arndt PF, Hwa T, Petrov DA (2006) A novel method distinguishes between mutation rates and fixation biases in patterns of single-nucleotide substitution. J Mol Evol 62: 168-175

Liu Z, Hossain GS, Islas-Osuna MA, Mitchell DL, Mount DW (2000) Repair of UV damage in plants by nucleotide excision repair: Arabidopsis UVH1 DNA repair gene is a homolog of Saccharomyces cerevisiae Rad1. Plant J 21: 519–528

Ludovisi R, Tauro F, Salvati R, Khoury S, Mugnozza GS, Harfouche A (2017) UAV-based thermal imaging for high-throughput field phenotyping of black poplar response to drought. Front Plant Sci 8: 1681

Lynch M (2010) Rate, molecular spectrum, and consequences of human mutation. Proc Natl Acad Sci U S A 107: 961-968

Madec S, Baret F, De Solan B, Thomas S, Dutartre D, Jezequel S, Hemmerlé M, Colombeau G, Comar A (2017) High-throughput phenotyping of plant height: comparing unmanned aerial vehicles and ground lidar estimates. Front Plant Sci 8: 2002

Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A, et al (2016) The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature 538: 201–206

Marti J, Bort J, Slafer GA, Araus JL (2007) Can wheat yield be assessed by early measurements of normalized difference vegetation index? Ann Appl Biol 50: 253-257

Masialeti I, Egbert S, Wardlow B (2010) A comparative analysis of phenological curves for major crops in Kansas. GIScience Remote Sens 47: 241-259

Mathieson I, Reich D (2017) Differences in the rare variant spectrum among human populations. PLoS Genet 13: e1006581

Matzke MA, Mosher RA (2014) RNA-directed DNA methylation: An epigenetic pathway of increasing complexity. Nat Rev Genet 15: 394-408

Meyer RS, Duval AE, Jensen HR (2012) Patterns and processes in crop domestication: An historical review and quantitative analysis of 203 global food crops. New Phytol 196: 29-48

Meyer RS, Purugganan MD (2013) Evolution of crop species: genetics of domestication and diversification. Nat Rev Genet 14: 840–852

Mirik M, Jones DC, Price JA, Workneh F, Ansley RJ, Rush CM (2011) Satellite remote sensing of wheat infected by wheat streak mosaic virus. Plant Dis 95: 4-12

Mitchell D, Bridge R (2006) A test of Chargaff's second rule. Biochem Biophys Res Commun 340: 90-94

Mooers A, Holmes EC (2000) The evolution of base composition and phylogenetic inference. Trends Ecol Evol 15: 365-369

Moss RA, Loomis WE (1952) Absorption spectra of leaves. I. the visible spectrum. Plant Physiol 27: 370-391

Muto A, Osawa S (1987) The guanine and cytosine content of genomic DNA and bacterial evolution. Proc Natl Acad Sci U S A 84: 166-169

Nasir AK, Tharani M (2017) Use of Greendrone UAS system for maize crop monitoring. Int Arch Photogramm Remote Sens Spat Inf Sci - ISPRS Arch 42: 263-268

Nawkar GM, Maibam P, Park JH, Sahi VP, Lee SY, Kang CH (2013) UV-Induced cell death in plants. Int J Mol Sci 14: 1608–1628

Odell ID, Wallace SS, Pederson DS (2013) Rules of engagement for base excision repair in chromatin. J Cell Physiol 228: 258-266

Olsen KM, Wendel JF (2013) A bountiful harvest: genomic insights into crop domestication phenotypes. Annu Rev Plant Biol 64: 47–70

Øvergaard SI, Isaksson T, Kvaal K, Korsaeth A (2010) Comparisons of two hand-held, multispectral field radiometers and a hyperspectral airborne imager in terms of predicting spring wheat grain yield and quality by means of powered partial least squares regression. J Near Infrared Spectrosc 18: 247–261

Pan X, Ma L, Zhang Y, Wang J, Du J, Guo X (2017) Three-dimensional reconstruction of maize roots and quantitative analysis of metaxylem vessels based on X-ray micro-computed tomography. Can J Plant Sci 98: 457-466

Panier S, Boulton SJ (2014) Double-strand break repair: 53BP1 comes into focus. Nat Rev Mol Cell Biol 15: 7-18

Peña JM, Torres-Sánchez J, de Castro AI, Kelly M, López-Granados F (2013) Weed mapping in early-season maize fields using object-based analysis of unmanned aerial vehicle (UAV) images. PLoS One 8: e77151

Purugganan MD, Fuller DQ (2009) The nature of selection during plant domestication. Nature 457: 843–848

Raun WR, Solie JB, Johnson G V., Stone ML, Lukina E V., Thomason WE, Schepers JS (2001) In-season prediction of potential grain yield in winter wheat using canopy reflectance. Agron J 93: 131-138

Ray DK, Mueller ND, West PC, Foley JA (2013) Yield trends are insufficient to double global crop production by 2050. PLoS One 8: e66428

Rist F, Herzog K, Mack J, Richter R, Steinhage V, Töpfer R (2018) High-precision phenotyping of grape bunch architecture using fast 3D sensor and automation. Sensors (Basel) 18: 763

Robert PC, Rust RH, Larson WE, Zhang M, Hendley P, Drost D, Robert PC, Rust RH, Larson WE, O'Neill M, et al (1999) Corn and soybean yield indicators using remotely sensed vegetation index. Precision Agriculture.ASA, CSSA, SSSA, Madison, WI. p.1475-1481

Robertson KD (2005) DNA methylation and human disease. Nat Rev Genet 6: 597-610

Romiguier J, Ranwez V, Douzery EJP, Galtier N (2010) Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. Genome Res 20: 1001-1009

Ross-Ibarra J, Morrell PL, Gaut BS (2007) Plant domestication, a unique opportunity to identify the genetic basis of adaptation. Proc Natl Acad Sci U S A 104 Suppl 1: 8641–8648

Royo C, Aparicio N, Villegas D, Casadesus J, Monneveux P, Araus JL (2003) Usefulness of spectral reflectance indices as durum wheat yield predictors under contrasting Mediterranean conditions. Int J Remote Sens 24: 4403-4419

Rudner R, Karkas JD, Chargaff E (1968) Separation of B. subtilis DNA into complementary strands. 3. direct analysis. Proc Natl Acad Sci U S A 60: 921-922

Sancar A (2003) Structure and function of DNA photolyase and cryptochrome blue-light photoreceptors. Chem Rev 103: 2203-2237

Schärer OD (2003) Chemistry and biology of DNA repair. Angew Chem Int Ed Engl 42: 2946-2974

Secco D, Wang C, Shou H, Schultz MD, Chiarenza S, Nussaume L, Ecker JR, Whelan J, Lister R (2015) Stress induced gene expression drives transient DNA methylation changes at adjacent repetitive elements. Elife 4: e09343

Serrano L, Filella I, Peñuelas J (2000) Remote sensing of biomass and yield of winter wheat under different nitrogen supplies. Crop Sci 40: 723-731

Shanahan JF, Schepers JS, Francis DD, Varvel GE, Wilhelm WW, Tringe JM, Schlemmer MR, Major DJ (2001) Use of remote-sensing imagery to estimate corn grain yield. Agron J 93: 583-589

Sharma S, Javadekar SM, Pandey M, Srivastava M, Kumari R, Raghavan SC (2015) Homology and enzymatic requirements of microhomology-dependent alternative end joining. Cell Death Dis 6: e1697

Sharp PM, Matassi G (1994) Codon usage and genome evolution. Curr Opin Genet Dev 4: 851–860

Shi Y, Thomasson JA, Murray SC, Pugh NA, Rooney WL, Shafian S, Rajan N, Rouze G, Morgan CLS, Neely HL, et al (2016) Unmanned aerial vehicles for high-throughput phenotyping and agronomic research. PLoS One 11: e0159781

Slotkin RK, Martienssen R (2007) Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet 11: e0159781

Šmarda P, Bureš P (2012) The Variation of Base Composition in Plant Genomes. Plant Genome Divers Vol 1 Springer, Vienna.

Song QX, Lu X, Li QT, Chen H, Hu XY, Ma B, Zhang WK, Chen SY, Zhang JS (2013) Genome-wide analysis of DNA methylation in soybean. Mol Plant 6: 1961–1974

Spitkó T, Nagy Z, Zsubori ZT, Szőke C, Berzy T, Pintér J, Marton CL (2016) Connection between normalized difference vegetation index and yield in maize. Plant, Soil Environ 7: 293-298

Springer NM, Schmitz RJ (2017) Exploiting induced and natural epigenetic variation for crop improvement. Nat Rev Genet 18: 563–575

Sueoka N (1962) On the genetic basis of variation and heterogeneity of DNA base composition. Proc Natl Acad Sci U S A 48: 582-592

Sumner AT, de la Torre J, Stuppia L (1993) The distribution of genes on chromosomes: A cytological approach. J Mol Evol 37: 117-122

Takahashi S, Sakamoto A, Sato S, Kato T, Tabata S, Tanaka A (2005) Roles of Arabidopsis AtREV1 and AtREV7 in translesion synthesis. Plant Physiol 138: 870–881

Takuno S, Gaut BS (2012) Body-methylated genes in Arabidopsis thaliana are functionally important and evolve slowly. Mol Biol Evol 29: 219-227

Tattaris M, Reynolds MP, Chapman SC (2016) A direct comparison of remote sensing approaches for high-throughput phenotyping in plant breeding. Front Plant Sci 7: 1131

Tenzen T, Yamagata T, Fukagawa T, Sugaya K, Ando A, Inoko H, Gojobori T, Fujiyama A, Okumura K, Ikemura T (1997) Precise switching of DNA replication timing in the GC content transition area in the human major histocompatibility complex. Mol Cell Biol 17: 4043-4050

Tommasi S, Denissenko MF, Pfeifer GP (1997) Sunlight induces pyrimidine dimers preferentially at 5-methylcytosine bases. Cancer Res 57: 4727–4730

Vergara-Díaz O, Zaman-Allah MA, Masuka B, Hornero A, Zarco-Tejada P, Prasanna BM, Cairns JE, Araus JL (2016) A novel remote sensing approach for prediction of maize yield under different conditions of nitrogen fertilization. Front Plant Sci 7: 666

Virlet N, Sabermanesh K, Sadeghi-Tehran P, Hawkesford MJ (2017) Field Scanalyzer: An automated robotic field phenotyping platform for detailed crop monitoring. Funct Plant Biol 44: 143-153

Wahab I, Hall O, Jirström M (2018) Remote sensing of yields: application of UAV imagery-derived NDVI for estimating maize vigor and yields in complex farming systems in sub-Saharan Africa. Drones 2: 28

Walter A, Liebisch F, Hund A (2015) Plant phenotyping: from bean weighing to image analysis. Plant Methods 11: 14

Wang JC (2002) Cellular roles of DNA topoisomerases: a molecular perspective. Nat Rev Mol Cell Biol 3: 430-40

Wang R, Cherkauer K, Bowling L (2016) Corn response to climate stress detected with satellite-based NDVI time series. Remote Sens 8: 269

Watanabe K, Guo W, Arai K, Takanashi H, Kajiya-Kanegae H, Kobayashi M, Yano K, Tokunaga T, Fujiwara T, Tsutsumi N, et al (2017) High-throughput phenotyping of sorghum plant height using an unmanned aerial vehicle and its application to genomic prediction modeling. Front Plant Sci 8: 421

Waters LS, Minesinger BK, Wiltrout ME, D'Souza S, Woodruff R V., Walker GC (2009) Eukaryotic translesion polymerases and their roles and regulation in DNA damage tolerance. Microbiol Mol Biol Rev 73: 134-154

Weber VS, Araus JL, Cairns JE, Sanchez C, Melchinger AE, Orsini E (2012) Prediction of grain yield using reflectance spectra of canopy and leaves in maize plants grown under different water regimes. F Crop Res 128: 82–90

West PT, Li Q, Ji L, Eichten SR, Song J, Vaughn MW, Schmitz RJ, Springer NM (2014) Genomic distribution of H3K9me2 and DNA methylation in a maize genome. PLoS One 9: e105267

White JW, Andrade-Sanchez P, Gore MA, Bronson KF, Coffelt TA, Conley MM, Feldmann KA, French AN, Heun JT, Hunsaker DJ, et al (2012) Field-based phenomics for plant genetics research. F Crop Res 133: 101-112

Wolfe KH, Sharp PM, Li WH (1989) Mutation rates differ among regions of the mammalian genome. Nature 337: 283-285

Yang GJ, Liu JG, Zhao CJ, Li ZH, Huang YB, Yu HY, Xu B, Yang XD, Zhu DM, Zhang XY, et al (2017) Unmanned aerial vehicle remote sensing for field-based crop phenotyping: current status and perspectives. Front Plant Sci 8: 1111

Yi C, He C (2013) DNA repair by reversal of DNA damage. Cold Spring Harb Perspect Biol 5: a012575

Zhang H, Lang Z, Zhu JK (2018) Dynamics and function of DNA methylation in plants. Nat Rev Mol Cell Biol 19: 489-506

Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SWL, Chen H, Henderson IR, Shinn P, Pellegrini M, Jacobsen SE, et al (2006) Genome-wide high-resolution mapping and functional analysis of DNA methylation in Arabidopsis. Cell 126: 1189-1201

CHAPTER 2. GENOME-WIDE NUCLEOTIDE PATTERNS AND POTENTIAL MECHANISMS OF GENOME DIVERGENCE FOLLOWING DOMESTICATION IN MAIZE AND SOYBEAN

Jinyu Wang[1], Xianran Li[1]*, Kyung Do Kim[2], Michael J. Scanlon[3], Scott A. Jackson[2], Nathan M. Springer[4], Jianming Yu[1]*

[1]Department of Agronomy, Iowa State University, Ames, IA 50011, USA
[2]Center for Applied Genetic Technologies, University of Georgia, Athens, GA 30602, USA
[3]Plant Biology Section, School of Integrative Plant Science, Cornell University, Ithaca, NY 14853, USA
[4]Department of Plant and Microbial Biology, University of Minnesota, St. Paul, MN 55108, USA
*Correspondence should be addressed to J.Y. (jmyu@iastate.edu), or X.L. (lixr@iastate.state)

Modified from a paper published in Genome Biology

Abstract

Plant domestication provides a unique model to study genome evolution. Many studies have been conducted to examine genes, genetic diversity, genome structure, and epigenome changes associated with domestication. Interestingly, domesticated accessions have significantly higher [A] and [T] values across genome-wide polymorphic sites than accessions sampled from the corresponding progenitor species. However, the relative contributions of different genomic regions to this genome divergence pattern and underlying mechanisms have not been well characterized. Here, we investigate the genome-wide base composition patterns by analyzing millions of SNPs segregating among 100 accessions from a teosinte-maize comparison set and among 302 accessions from a wild-domesticated soybean comparison set. We show that the non-genic part of the genome has a greater contribution than genic SNPs to the [AT]-increase observed between wild and domesticated accessions in maize and soybean. The separation between wild and domesticated accessions in [AT] values is significantly enlarged in non-genic and pericentromeric regions. Motif frequency and sequence context analyses show the motifs (PyCG) related to solar-UV signature are enriched in these regions, particularly when they are

methylated. Additional analysis using population-private SNPs also implicates the role of these motifs in relatively recent mutations. With base-composition across polymorphic sites as a genome phenotype, genome scans identify a set of putative candidate genes involved in UV damage repair pathways. The [AT]-increase is more pronounced in genomic regions that are non-genic, pericentromeric, transposable elements, methylated, and with low recombination. Our findings establish important links among UV radiation, mutation, DNA repair, methylation, and genome evolution.

Introduction

Domestication is a special mode of evolution. Extensive studies have been carried out to understand domestication process and genes associated with morphological changes [1-4]. Meanwhile, genomes also went through profound changes during domestication. Recent studies documented the base-composition difference and mutation rate difference between populations separated by either a domestication or demographic bottleneck event, which provide novel insights on genome evolution [5-7]. Further investigation in DNA base composition, mutation spectrum, and the potential relationship between them are necessary to advance our understanding of genome changes.

DNA base composition is an essential genomic feature. Remarkable research progress has been made in several areas, including codon usage bias [8], isochore structure [9, 10], and GC-biased gene conversion [11]. Recently, a conserved base-composition pattern, modern accessions having significantly higher [A] and [T] values across genome-wide polymorphic sites than accessions sampled from their wild relatives, was discovered with natural populations across multiple species [5]. Different genomic regions exhibit different patterns of a number of genomic features such as DNA methylation, GC content, and recombination rate [12-15]. It

would be interesting to study the regional variation of genome change pattern, captured by base-composition summarized from polymorphic sites.

Mutation is a fundamental factor that generates the genetic variation upon which selection, drift, and recombination act. Point mutations are the most common type of mutations with a universal bias towards high AT, primarily due to the high rate of transition mutations [16]. Recent studies indicated that mutation rate can be different across populations [6, 7]. Divergence in mutation rates or types between populations are one of several factors that affect genetic variation patterns [17]. Analysis of data from multiple mutation accumulation experiments, either accumulating spontaneous or induced mutations, demonstrated that higher [AT] values across mutation sites in derived lines at the end of mutation experiments than in ancestral lines, which suggested that base-composition difference can emerge from mutation sites [5]. Characterization of mutation spectrum in natural populations may help unravel the mechanism of genome change [18].

Organisms have evolved a complex system to monitor and repair DNA damage caused by various exogenous mutagens, such as solar-ultraviolet (UV) radiation, reactive oxygen species, excess boron or aluminum, and pathogenic microorganisms [19]. For plants, solar-UV radiation is a major exogenous mutagen as they use sun light for photosynthesis. The primary solar UV-induced DNA lesion, cyclobutane pyrimidine dimers (CPDs), induces C→T base transitions [20]. CPDs distort the DNA's double-helix structure, which influences DNA unwinding and DNA replication, and ultimately affect cell cycle [21]. Using sets of SNPs private to different human populations, a recent study suggested that UV might have been involved in the mutation spectrum change [6].

DNA methylation is a major form of epigenetic modification in many eukaryotic genomes. It not only regulates gene expression and silences transposons and repeat sequences,

but also affects mutation rates [22-25]. DNA methylation occurs in CG, CHG (where H = A, C, or T) and CHH sequence contexts in plants [26, 27]. The relative frequency of DNA methylation varies substantially along chromosome. DNA methylation is primarily distributed in the heterochromatin regions that are mostly composed of tandem repeats and transposons [12, 13, 28]. It has been shown that methylation of cytosine residue at CpG sites can enhance the solar UV-promoted CPD formation [25]. We can ask whether the rate of solar-UV induced mutations varies along the chromosome and whether base-composition can summarize such variation.

In this study, we report findings from the analysis of millions of SNPs segregating among 100 accessions from a teosinte-maize comparison set and among 302 accessions from a wild-domesticated soybean comparison set. First, we show that higher [AT] values in domesticated accessions relative to wild accessions, or [AT]-increase, is consistently observed for SNPs found in either genic or non-genic portions of the genome, with non-genic SNPs having a greater contribution to the [AT]-increase. Interestingly, we also find that the divergence in [AT] is much higher in pericentromeric regions than other regions. All four sequence motifs related to solar-UV signature consistently have higher frequencies in methylated regions than unmethylated regions. With a different set of population-private SNPs, we also discover the enrichment of mutations related to the solar-UV signature in domesticated accessions. Using base-composition across polymorphic sites as the phenotype, genome-wide scans identify a set of putative candidate genes involved in UV damage repair pathways. Together, these findings seem to suggest that solar-UV radiation and differential mutation repair are critical components in the genome divergence process that resulted in domesticated accessions' greater numbers of nucleotide A and T.

Materials and Methods

Sequence information and SNP extraction

In maize, the original SNP set with B73 genome (AGPv2) as references was obtained from 103 maize genomes of Maize Hapmap2 (19 wild accessions, 23 landraces, and 61 improved cultivars) [29]. Three lines, 2 wild accessions and 1 improved cultivar, were removed due to low sequence coverage and a small number of SNPs. In soybean, the original SNP set with Williams 82 genome (version 1.1) as references was obtained from 302 soybean genomes (62 wild accessions, 130 landraces, and 110 improved cultivars) [30]. Information for maize and soybean accessions are provided in Table S5 and Table S6, respectively. With CrossMap v0.2.5 [84], genome coordinates of the original SNP sets in B73 AGPv2 and Williams 82 version 1.1 were converted to that in B73 AGPv4 and Williams 82 version 2.0, respectively. In maize, the assembly chain file for CrossMap is available at ftp://ftp.ensemblgenomes.org/pub/plants/release-39/assembly_chain/zea_mays/AGPv2_to_AGPv4.chain.gz. And in soybean, the assembly chain file is available at ftp://ftp.ensemblgenomes.org/pub/plants/release-39/assembly_chain/glycine_max/V1.0_to_Glycine_max_v2.0.chain.gz.

Then for each species, we obtained two sets of SNPs (common SNP set and population-private SNP set) from the original SNP sets by applying different filtering criteria (Figure S1). The common SNP sets containing 8,852,678 SNPs in maize and 4,870,265 in soybean are obtained by filtering with a MAF threshold of 5% and a missing rate threshold of 20%. These common SNP sets are used for all analyses except population-private SNP analysis.

For population-private SNP sets, we followed the procedure laid out in a previous study [6] to obtain 2,651,790 population-private SNPs in maize and 681,791 population-private SNPs in soybean. The private SNP sets are different from the common SNP sets with a small overlap.

Ancestral state of the maize allele was inferred based on the allele of *Tripsacum* [49]. To infer the ancestral state of the soybean allele, BLASTN [85] (version 2.2.28+) was used to identify the orthologous regions between soybean and *Medicago truncatula*. Each SNP and its 58 bases flanking sequences were extracted from soybean genome, then blasted to the *Medicago truncatula* genome sequence [86] with an e-value < $1e^{-1}$ and only the best hit was considered. A SNP is considered as population-private if it is segregating in one group but fixed ancestral allele in other groups. Based on this definition, we obtained 1,137,732 private wild SNPs (PW) that are segregating in wild group but fixed ancestral allele in landrace and improved cultivar group, 1,514,058 private domesticated SNPs (PD) that are segregating in either landrace or improved group but fixed ancestral allele in wild group, 270,390 private landrace SNPs (PL) that are segregating in landrace group but fixed ancestral allele in wild and improved cultivar group, and 537,259 private improved cultivar SNPs (PI) that are segregating in improved cultivar group but fixed ancestral allele in wild and landrace group. In soybean, we obtained 571,756 PW, 110,035 PD, 20,543 PL, and 1,798 PI. The total numbers of SNPs (2,651,790 in maize and 681,791 in soybean) in private SNP sets are obtained by summing up PW and PD because there are no overlapping SNPs between the two population-private SNP sets by definition.

For maize, all analyses were done using maize B73 genome (version AGPv4) as references. For soybean, all analyses were done using soybean Williams 82 genome (version 2.0) as references. *Medicago truncatula* genome sequence (version Mt4.0) was downloaded from Phytozome. Short reads from representative soybean accessions were downloaded from GenBank.

Bioinformatics

DNA reads were mapped to soybean reference genome by BWA with the BWA-MEM algorithm [87]. R packages *Rsamtools* [88] and *GenomeGraphs* [89] were used to analyze and

display the sequence coverage in candidate genes. The missing genotypes in candidate genes were imputed by fastPhase under the context including up- and down-stream 20kb regions [90]. R package *pegas* was used to reconstruct the haplotype networks with SNPs detected in the genes [91]. All the other analyses are done with in-house scripts written in Perl or R. Base-composition across genome-wide SNP sites was calculated as described in previous study [5]. Because of PR2, *i.e.*, nucleotide A content ([A]) from SNP sites is roughly equals to [T] ([A] ≈ [T]) and [C] ≈ [G] [5], the value of [AT] was used in this study.

Base composition distribution among substitution types

Bi-allelic SNPs can be grouped into 6 substitution types (A/C, A/G, A/T, C/G, C/T, and G/T) without defined ancestral allele. For example, if C and T alleles are detected in one SNP site, which might arise either from C to T change or from T to C change, it is a C/T substitution type. For each substitution type, the total number of each nucleotide type possessed by each accession was counted and divided by the total number of polymorphic sites (8.9 million in maize and 4.9 million in soybean for the accession without missing calls).

Base composition distribution at different genomic regions

SNP effects were predicted with the SnpEff v4.3 [92]. In maize, we built the database with reference genome sequences available at ftp://ftp.ensemblgenomes.org/pub/plants/release-39/fasta/zea_mays/dna/Zea_mays.AGPv4.dna.toplevel.fa.gz and gene annotation available at ftp://ftp.ensemblgenomes.org/pub/plants/release-39/gff3/zea_mays/Zea_mays.AGPv4.39.chr.gff3.gz. In soybean, we built the database with reference genome sequences available at ftp://ftp.ensemblgenomes.org/pub/plants/release-39/fasta/glycine_max/dna/Glycine_max.Glycine_max_v2.0.dna.toplevel.fa.gz and gene annotation available at ftp://ftp.ensemblgenomes.org/pub/plants/release-39/gff3/glycine_max/Glycine_max.Glycine_max_v2.0.39.chr.gff3.gz.

Seven genomic annotation sets (intergenic, gene-proximal, UTRs, synonymous, missense, intronic and other genic) were obtained by classifying SNPs based on the predicted SNP effect. SNPs were classified to be gene-proximal if they fell within 5 kb upstream of the transcription start site. Then intergenic set together with gene-proximal set is considered as non-genic SNP set, and the rest five SNP sets are considered to be genic SNP set. After that, base-composition across polymorphic sites was calculated for genic SNP set and non-genic SNP set separately.

The physical positions for maize centromeric corresponding to genome (version AGPv4) were referred from a previous study [93]. Then a 40 Mb segment directly adjacent upstream and downstream of the centromeric region were considered as pericentromeric regions based on a previous study [33]. And the physical coordinates for soybean centromeric and pericentromeric regions were obtained from [34] and Soybean Genome Browser at SoyBase https://soybase.org/gb2/gbrowse/gmax2.0/.

To analyze the base-composition distribution along chromosomes, we calculated the [AT] for each accession with a moving average approach of a 5-Mb window size and a 4-Mb step size on each of the maize and soybean chromosomes with both genic and non-genic SNPs. Indeed, we examined the [AT] distribution with a series of window size including 1-Mb, 2-Mb, 5-Mb, and 10-MB. The patterns for all of those window sizes are similar. We decided to go with the 5-Mb for the analyses because it contains a good amount of SNPs in each window and the line of [AT] distribution is smoother than the smaller window size.

Position of crossovers (COs) in maize were referred from [39]. Then [AT]-difference and crossover (CO) rate were calculated using a 5-Mb sliding window. Recombination rate data in soybean was referred from [30]. [AT]-difference and recombination rate were calculated using a

1-Mb window. Correlation was calculated between [AT]-difference and CO rate or recombination rate for each chromosome.

Transposable element (TE) regions in maize and soybean are referred from [93, 94]. Then base-composition across polymorphic sites was calculated for SNPs within TE regions and non-TE regions separately.

Selective sweep regions in maize and soybean are referred from [29, 30]. Then base-composition across polymorphic sites was calculated for SNPs within selective sweep and non-selective-sweep regions separately.

The maize methylation data was generated from whole-genome bisulfite sequencing (WGBS) of leaf tissue of maize B73 seedling [42]. Genome coordinates of B73 methylation data in AGPv2 were converted to that in AGPv4 with the CrossMap v0.2.5 [84]. Then the maize genome was separated into methylated and unmethylated regions based on whether the percentage of CG methylation within each 100bp non-overlapping window is greater than 40% or not. The soybean methylation data was generated from WGBS of leaf of soybean Williams 82 [43] and GsojaD [44].

MethylC-seq reads of GsojaD were first mapped to its own genome assembly to get methylation call. Then the genome coordinates of GsojaD methylation were converted to the coordinates in Williams 82 genome version 2. Genome coordinates of Williams 82 methylation data in Williams 82 version 1.1 were converted to those in version 2.0 with the CrossMap v0.2.5 [84]. Then the soybean genome was separated into the methylated and unmethylated regions based on CG methylation sites that are common to both Williams 82 and GsojaD.

Motif enrichment analysis

For each SNP site, the directly adjacent upstream and downstream bases were extracted from reference genomes, meanwhile, the adjacent sequences of one randomly selected site from

1kb flanking region were also extracted. For each of the 96 possible tri-nucleotide motifs (5′-NXN-3′, X is the polymorphic site or randomly selected site), an empirical threshold at 95th percentile was drawn from 100 random sample scenarios. A motif is considered as enriched if the ratio of its frequency at SNP site over the 95th percentile at random site is greater than 1.

Population-private SNP analysis

We used the procedure laid out in a previous study [6] to test mutation spectrum differences between populations with population-private SNPs. SNPs within each private SNP set were partitioned into 96 mutation types through considering the base immediately upstream and downstream of the variable site [47]. Count data $C_p(m)$ of type-m mutations in set P for each mutation type m = $B_{5'}B_AB_{3'} \rightarrow B_{5'}B_DB_{3'}$ of each private SNP set P were obtained. Then with a χ^2 test, $f_{PI}(m)$ and $f_{PL}(m)$ were compared with $f_{PW}(m)$. For the χ^2 test, we used χ^2 value instead of P-value to indicate the significance of difference because P-value cannot be obtained for very large χ^2 value in our data.

To assess the variance of $f(TCG \rightarrow \mathrm{T})$ and $f(CCG \rightarrow \mathrm{T})$, private SNP sets PL, PI and PW in maize and PD and PW in soybean was partitioned into non-overlapping bins of 1,000 consecutive SNPs. Then $f(TCG \rightarrow \mathrm{T})$ and $f(CCG \rightarrow \mathrm{T})$ for each bin were calculated.

GWAS for base composition in maize and soybean

Following our earlier study in human [5], [AT] values across 8,852,678 maize SNPs and 4,870,265 soybean SNPs were used as the genome phenotype for GWAS. In the genome scan for both maize and soybean, a mixed linear model (MLM) with both fixed covariates and a random kinship matrix was used to detect SNPs associated with the base composition variation [95, 96] in GAPIT version 3.35 [97]. Parameters in MLM were determined by model selection process [95, 96]. Five principal components (PC2-PC6) were selected in maize and 0 PC was selected in

soybean. PC1 was not under model selection process because of its near perfect correlation with [AT] [5]. The significance threshold P-value was determined by Bonferroni correction.

The 334 maize genes and 107 soybean genes associated with repairing UV damaged DNA were compiled based on either the sequence similarity of rice genes or *Arabidopsis* genes [48]. We conducted enrichment test of UV-related genes with a series of window sizes centered by significantly associated SNPs as described in a previous study [31]. The proportion of UV-related genes within each window was compared with its genome-wide proportion. The gene was counted when it was tagged by at least 2 significantly associated SNPs. Then we tested whether the proportion of UV-related genes within the window is significantly higher than that across the whole genome using proportion test. The window size smaller than 500kb in maize and 200kb in soybean were not tested because their numbers of tagged UV-related genes were less than 10, which violated the condition of the proportion test.

Results

Genome-wide [AT]-increase

We obtained a set of 8,852,678 SNPs in 100 teosinte-maize accessions and a set of 4,870,265 SNPs in 302 wild-domesticated soybean accessions from the original studies [29, 30] (Figure S1). These SNPs are designated as common SNP sets to compute the genome-wide base composition across polymorphic sites without concerning about sampling issues due to low minor allele frequency (MAF) or high missing rate [5]. For each accession, we obtained a [AT] value calculated as the fraction of SNP alleles that are either base A or T. The choice of [AT] was based on the finding that single strand parity rule 2 (PR2) applies to base composition across SNPs [5], *i.e*. [A] ≈ [T] and [G] ≈ [C]. In both maize and soybean sets, wild and domesticated (including landraces and improved cultivars) accessions are clearly separated by [AT] (P-value is 1.49e-14 for maize and 1.02e-44 for soybean). Domesticated accessions have more nucleotide A

and T at the polymorphic sites (Figure 1), termed as [AT]-increase (domesticated > wild accessions). In maize, the average value of [AT] in wild accessions is 0.380 (SD=0.006), while the average values of [AT] in landraces and improved cultivars are 0.414 (SD=0.003) and 0.417 (SD=0.003), respectively. In soybean, the average value of [AT] in wild accessions is 0.449 (SD=0.010), while the average values of [AT] in landraces and improved cultivars are 0.492 (SD=0.006) and 0.494 (SD=0.003), respectively.

Base-composition among DNA substitution types

Bi-allelic SNPs can be grouped into 6 substitution types without defining the ancestral allele. To further understand the consistent [AT]-increase pattern, we examined the contribution to [AT]-increase from each substitution type (Figure S2). Two transition types, A/G and C/T, are the major types detected in maize and soybean genomes, with each having a frequency of ~34%, much higher than the expected frequency by chance (*i.e.*, ~17% or 1/6). The average frequency for each of four transversion types (A/C, A/T, C/G, and G/T), is less than 10%, with C/G type being the least frequent one.

We then calculated base-composition value across polymorphic sites conditional on each substitution type. The contribution to the overall [AT]-increase varied among substitution types (Figure 2). Two transition types (A/G and C/T) are the major contributors due to their high frequencies and that the majority of wild accessions possess G or C allele for these types, while the domesticated typically have A or T. For A/C and G/T types, significant base-composition differences between wild and domesticated groups are also evident, and the proportional increase in A or T is similar to that of A/G and C/T types. However, because of their relatively low frequencies (≤ 9%), these two types contribute less to the overall [AT]-increase. Neither A/T nor C/G type contributes to the overall [AT]-increase.

Base-composition pattern at different genomic regions

It is known that different genomic regions exhibit different patterns for a number of genomic features including DNA methylation, GC content, and recombination rate [12-15], which naturally led us to investigate the base-composition distribution at different parts of the genome. To facilitate this, we first classified the genome-wide SNPs to 7 genomic annotation sets: intergenic, gene-proximal, UTRs, synonymous, missense, intronic and other genic [31, 32] (Figure 3). Intergenic SNPs are the most common group (65.1% in maize and 57.4% in soybean), followed by gene-proximal (15.3% in maize and 26.6% in soybean) and intronic (10.9% in maize and 8.98% in soybean). Because the numbers of SNPs were relatively too small in several genomic annotation sets, we combined intergenic and gene-proximal sets to form the non-genic SNP set, and combined the rest five original genomic annotation sets to form the genic SNP set. The non-genic set contains 7,120,981 SNPs in maize and 4,088,443 SNPs in soybean, and the genic set contains 1,731,687 SNPs in maize and 781,822 SNPs in soybean.

We calculated the [AT] value for each accession from genic and non-genic SNP sets separately. [AT] of domesticated accessions is consistently higher than that of wild accessions in both genic and non-genic SNPs (Figure 3). However, non-genic SNPs have greater contributions to the [AT]-increase, and the [AT]-difference between wild and domesticated accessions is about twice that of genic SNPs. Since the total number of non-genic SNPs are 4 to 5.5 times larger than genic SNPs, we randomly sampled an equal number of SNPs from genic and non-genic SNP sets to obtain the [AT] value for comparison. We obtained a consistent trend from 100 subsets, demonstrating that the greater contribution to the overall [AT]-increase from non-genic SNPs is not only because of its larger SNP number but also due to its higher proportional increase in [AT] than genic SNPs (Figure S3). As expected, further comparisons of [AT] distribution between missense, synonymous, and intergenic SNP sets (Figure S4A-B) show that while [AT]-

difference between wild and domesticated accessions from missense and synonymous SNP sets are similar to each other, both of them are smaller than intergenic SNP set. We also evaluated the impact of allele frequency on the different contributions from genic and non-genic SNPs. Compared with non-genic SNP set, genic SNP set generally has more SNPs with high MAF and fewer SNPs with low MAF (Figure S5), which may suggest that the genic region is more conserved than non-genic regions.

Both species are known to have low gene density in pericentromeric regions [29, 33, 34], so we examined the [AT] distribution with genic and non-genic SNPs along chromosomes (Figure 3, Figure S6-S8). Along each chromosome, *a)* higher [AT] in domesticated group than wild group is consistently observed for both genic and non-genic SNPs; *b)* [AT]-difference between domesticated and wild group for non-genic SNPs is generally larger than that for genic SNPs; and *c)* [AT] for each accession is higher for genic SNPs than non-genic SNPs. More interestingly, the divergence in [AT] is significantly enlarged in pericentromeric regions, especially for non-genic SNPs.

Because of the dramatic difference of [AT] distributions between pericentromeric regions and chromosome arms, we further compared the [AT] distribution between genic and non-genic regions conditional on pericentromeric regions and chromosome arms separately (Figure S4C-F). The [AT]-difference between wild and domesticated accessions at non-genic region is consistently about twice that of genic regions for both pericentromeric regions and chromosomal arms. And the [AT]-difference between wild and domesticated accessions at pericentromeric region is much larger than that of chromosome arms, which is true for both non-genic and genic SNPs.

We speculate the enlarged [AT]-difference in pericentromeric regions is associated with the fact that these regions mainly consist of repetitive sequences and transposable elements [33-

36] that are mostly arranged in heterochromatin [37], and generally have low recombination rates [30, 33, 34, 38]. To verify the speculation, we first examined the distribution of base-composition at transposable element (TE) and non-transposable element (non-TE) regions. The [AT]-differences at TE regions are much larger than non-TE regions (Figure S9). We then plotted the [AT]-difference and crossover rate for maize and recombination rate for soybean along each chromosome (Figure S10-S12). Negative correlations between [AT]-difference and crossover/recombination rate are significant for all 10 maize chromosomes and 18 soybean chromosomes. We observed relatively low and fluctuating MAF within pericentromeric regions (Figure S13-S15), which may be related to the low efficiencies in purging out deleterious alleles [39].

As the phenotypic differences between the wild and domesticated accessions mainly shaped by the artificial selection, we then compared the base-composition distribution at domestication selective sweep and non-selective-sweep regions to test if the domestication process was partially responsible for the detected base-composition difference. The [AT]-difference between wild and domesticated accessions at selective sweep regions is much larger than that at non-selective-sweep regions (Figure S16). This suggests that the domestication process indeed have effect on the detected base composition difference at the polymorphic sites.

Enrichment of motifs related to solar-UV signature surrounding SNP sites

To test whether SNPs occurred more frequently in certain sequence contexts, we first classified SNPs into 96 tri-nucleotide motifs by considering one base directly adjacent upstream and downstream of the SNP site. Then we examined the frequency and the enrichment of tri-nucleotide motifs. With 96 possible motifs, the expected frequency is 0.010 (≈1/96) and a ratio of 1.000 between the frequency of motif at SNP sites and that at random sites if SNPs occurred randomly in the genome. We detected 14 common motifs between maize and soybean with both

frequencies and ratios greater than the expected, and 11 out of 14 were from A/G and C/T transition types (Figure 4). In both species, 5′-CNG-3′ (N is the polymorphic site) around C/T type has the highest ratio with 2.007 in maize and 2.228 in soybean. In addition, 5′-TNG-3′ is enriched around C/T type, with a ratio of 1.477 in maize and 1.311 in soybean. Because most wild accessions have C allele at C/T type (Figure 2), these SNPs were more likely changed from 5′-PyCG-3′ to 5′-PyTG-3′, where Py is either pyrimidine C or T. Correspondingly, the reverse and complementary motifs 5′-CNG-3′ and 5′-CNA-3′ around A/G type are also overrepresented, which suggests the high chance of 5′-CGPu-3′ to 5′-CAPu-3′ mutations, where Pu is purine G or A.

Solar UV induces CPDs preferentially at 5-methylcytosine-containing dipyrimidine sites (5′-Py-$_m$CG-3′), termed as solar-UV signature [20, 40]. Thus, the overrepresented motif 5′-PyCG-3′ around C/T (the reverse and complementary motif 5′-CGPu-3′ around A/G) is the same as solar-UV signature if C is methylated. Hereafter, we refer to the four aforementioned sequence motifs as motifs related to solar-UV signature. In both species, $_m$CG level is negatively correlated with gene density and enriched in pericentromeric regions [12, 13, 28, 41], which suggests that motifs related to solar-UV signature might occur more frequently outside of genic regions and be overrepresented in pericentromeric regions. To test this, we performed two sets of comparisons: frequencies of motifs related to solar-UV signature between genic and non-genic SNPs, and between SNPs from pericentromeric and non-pericentromeric regions. As expected, all four motifs related to solar-UV signature have higher frequencies within non-genic SNP sets than genic SNP sets, and they have higher frequencies among SNPs from pericentromeric regions than among SNPs from non-pericentromeric regions (Figure S17-S18).

We then examined the role of DNA methylation by calculating the frequencies of motifs related to solar-UV signature conditional on methylated and unmethylated regions [42-44]. We

found that all four motifs related to solar-UV signature consistently have higher frequencies in methylated regions than unmethylated regions with genic SNPs, non-genic SNPs, SNPs from pericentromeric regions, and SNPs from non-pericentromeric regions (Figure S17-S18). This suggests the higher probability of C→T and G→A transitions, potentially stimulated by DNA methylation, in non-genic regions and pericentromeric regions, which agrees with our findings of non-genic SNPs' larger contributions to [AT]-difference and the enlarged [AT]-difference in pericentromeric regions.

Mutation spectra of population-private variation

The findings of sequence motifs related to solar-UV signature enriched in common SNP sets encourage us to verify the pattern with rare segregating SNPs that occurred as relatively recent mutations [45, 46]. Therefore, following procedures laid out in a previous study [6], we compiled private SNP sets that contain 2,651,790 population-private SNPs in maize and 681,791 population-private SNPs in soybean from original studies [29, 30] (Figure S1). These private SNP sets are different from the earlier common SNP sets with a small overlap. A SNP is considered as population private if it is segregating in one lineage but fixed ancestral allele in other lineages. For each crop, we obtained four population-private SNP sets: private wild SNPs (PW), private domesticated SNPs (PD), private landrace SNPs (PL) and private improved cultivar SNPs (PI). PW designates SNPs that are segregating in the wild group but are fixed ancestral alleles in the landrace and the improved cultivar groups, PL means those SNPs are segregating in the landrace group but are fixed ancestral alleles in the wild and the improved cultivar groups, and similarly for other private SNP sets. Analyzing such SNPs enables us to assess the mutation rate difference among different lineages after diverged from the most recent common ancestor.

Next, we tested differences in the spectrum of mutagenesis between populations with population-private variants as described in the previous study [6]. With ancestral allele information, population-private SNPs can be partitioned into 96 mutation types by considering the base immediately upstream and downstream of the variable site [47]. In both species, most C→T transitions have higher frequencies in PL and PI than in PW, which agrees with previous finding in a human study [6] (Figure 5). This observation suggests although genomes of domesticated and wild accessions were continuing to evolve after divergence, domesticated accessions might have higher C→T mutation rate. We observed higher rate of mutations related to solar-UV signature 5′-TCG-3′→5′-TTG-3′ and 5′-CCG-3′→5′-CTG-3′ (hereafter abbreviated as TCG→T and CCG→T) in domesticated accessions than wild accessions (Figure 5, Figure S19). For instance, in maize, TCG→T has frequency of 3.45% in PL and 3.55% in PI compared with 2.99% in PW. The higher frequencies of TCG→T and CCG→T in domesticated than wild accessions are consistent for all chromosomes (Figure S20). We further split each population-private SNP set to genic-private SNPs and non-genic-private SNPs, and pericentromeric-private SNPs and non-pericentromeric-private SNPs. As shown by Figure S21, in both species, the TCG→T and CCG→T mutations generally have higher frequencies with non-genic-private SNPs and pericentromeric-private SNPs.

This overrepresentation of mutations related to solar-UV signature found in the private SNP sets together with the enrichment of motifs related to solar-UV signature found in the common SNP sets suggest that solar UV is potentially one of the major forces driving the [AT]-increase pattern during domestication.

Overrepresentation of genes repairing UV damaged DNA near loci associated with genome divergence

With genome-wide association studies (GWAS), the previous study in human found the enrichment of DNA repair genes surrounding loci associated with genome divergence captured

by base-composition across polymorphic sites [5]. The enrichment of solar-UV-signature mutations in domesticated accessions suggests that solar-UV radiation plays an important role in driving the [AT]-increase pattern. Plant genomes encode a complex system to monitor and repair DNA damage. We assessed whether genes involved in UV damage repair pathways are enriched near loci associated with genome divergence for [AT].

Using the [AT] values obtained from the common SNP sets as a genome phenotype, GWAS identified a series of loci significantly associated with base-composition across polymorphic sites (Figure S22). Based on either the sequence similarity of rice genes or *Arabidopsis* genes [48], 334 maize and 107 soybean genes were compiled as related to UV damaged DNA repair (UV-related gene hereafter). Proportion tests indicate that the UV-related genes were more likely to reside nearby GWAS signals than by chance (Table S1-S4). In maize, for the 500kb segments around significantly associated SNPs, we identified 4.2% of UV-related genes, but these regions only encode 1.8% of all annotated genes. In soybean, for the 500kb segments around significantly associated SNPs, 20.6% of UV-related genes were identified, while only 13.8% of annotated genes were encoded in these regions. The tagged genes involved in all the steps for global genome nucleotide excision repair (NER) pathway to repair UV damage are shown in Figure S23.

We performed detailed analysis of several UV-related genes located near significant GWAS SNPs (Figure 6). A SNP located within maize *ATR* (*Zm00001d014813*) is significantly associated with base-composition across polymorphic sites. The *ATR* encodes a putative ATR protein which functions in a wide range of responses to DNA damage, including sensing and activating a cell cycle arrest in response to UV-B caused DNA damage [19]. We found 8 nonsynonymous variants located in *ATR* in this maize population. In soybean, a SNP located 11kb downstream of *Ligase1* (*Glyma.11g193100*, *Lig1*) on chromosome 11 is strongly

associated with [AT] variation. *Lig1* in soybean encodes a putative DNA ligase 1 protein which functions in sealing the nick of DNA at the last step of repairing process. Besides one nonsense and two nonsynonymous SNPs, we also detected a 1.8kb deletion at the 5th intron in wild soybean accessions (Figure S24). Soybean genome encodes two copies of *Lig1*, and we did not detect signals for *Lig1* on chromosome 12.

Both *ATR* and *Lig1* are located within selective sweep regions identified in previous studies [30, 49], which suggests the possibility that polymorphisms within *ATR* and *Lig1* went through domestication bottleneck. We then conducted haplotype network analysis of these two genes. There are two distinct clusters of haplotypes in both *ATR* and *Lig1* (Figure 6), one composed mostly of domesticated accession haplotypes and the other composed mostly of wild accession haplotypes. We refer to these clusters as domesticated cluster haplotype (DCH) and the wild cluster haplotype. In *ATR,* DCH is present in >98% of maize but <18% of teosinte; while in *Lig1,* DCH is present in >97% of domesticated soybean but <5% of wild soybean. Intriguingly, the major haplotype (haplotype2) in both genes are shared by most of domesticated accessions and a small number of wild accessions. Haplotype2 in *ATR* is shared among 86.7% of maize and 17.6% of teosinte, and haplotype2 in *Lig1* is shared by 86.7% of domesticated soybean and 2% of wild soybean. Considering that domestication largely involved selection of favorable alleles from standing allelic variation in wild ancestors [1], it is likely that the major haplotypes for both *ATR* and *Lig1* were present in the ancestral populations with low frequency, and their frequencies increased rapidly during domestication.

Discussion

Our understanding of how plant genomes have changed following domestication bottlenecks remains limited. In this study, we aim to address the question from a novel angle by surveying the genome-wide base composition pattern and its potential associated mechanisms.

Focusing on a genome phenotype summarized from millions of polymorphic sites along the chromosome, we provide novel insights on genome evolution at different parts of the genome: genic versus non-genic, pericentromeric versus non-pericentromeric, and methylated versus unmethylated. This study also presents a first case where a few critical components in genome evolution are brought together: “Base composition”, “Mutation”, “UV radiation”, “DNA repair”, and “Methylation”.

The [AT]-increase in domesticated over wild accessions is consistently observed with the overall genome-wide SNPs, SNPs within major genomic annotation sets, and SNPs from different genomic regions. These findings indicate the presence of common underlying mechanisms that drive the domesticated accessions to build their genomes with more A and T nucleotides. In both maize and soybean, the SNP sets were obtained by aligning to a reference genome from the domesticated group. We acknowledge mapping bias exists, but we do not expect resolving mapping bias by aligning to a reference genome from the ancestral group would change the observed [AT]-increase patterns.

Demographical analyses have shown that plant and animal species experienced population size changes associated with domestication and range expansion [50-54]. The effective population size of maize has decreased strikingly from the onset of domestication (≈ 10,000 years ago) to the recent past (≈ 1100 – 2400 years ago) and increased during post-domestication expansion [50]. In contrast to maize, the wild *parviglumis* experienced an increase in effective population size which also lasts until the recent past (≈ 1100 – 1800 years ago) [50]. In plants, increased mutational load has been observed in populations that undergo declines in effective population size [50, 55, 56]. Thus, one interpretation for our findings is that, domesticated populations have historically lower effective population size, which results in stronger genetic drift and relaxed purifying selection, and consequently lead to higher mutation

numbers compared with their wild relatives. Meanwhile, our discovery of the overrepresentation of mutations related to solar-UV signature in domesticated accessions indicated varied mutation rate across populations. Therefore, an alternative interpretation is that alleles of UV damage repair genes have different repair efficiency (lower in domesticated accessions) and affect the number of *de novo* mutations in different lineages.

Regarding the increased [AT] in domesticated accessions, one natural question to ask is: what is the consequence of building genomes with more A and T nucleotides? One possibility will be more efficient energy usage. Energy usage efficiency is a trait under universal selection that has shaped various genomic aspects. For example, highly expressed proteins use cheaper amino acids [57-60] and are generally shorter than lowly expressed ones [61, 62]. Synthesizing a G+C basepair requires larger amount of energy and nitrogen than producing an A+T basepair [63]. Base stacking for G and C is more energetically expensive compared with that for A and T, as G binds to C with three hydrogen bonds while A binds to T with two hydrogen bonds [64]. Therefore, it may be interesting to ask whether domesticated accessions build their genomes with more A and T so that more energy is saved for other biological processes toward better yield potential.

Recent studies have showed the high heterogeneity of mutation rate across genomic regions [65-67]. Our survey discovered the enrichment of motifs related to solar-UV signature surrounding SNPs, especially for SNPs located in non-genic and pericentromeric regions, which suggests solar-UV radiation is likely one of the major contributors for plant genome divergence. In general, DNA methylation level of non-genic regions is higher than genic regions, and pericentromeric regions higher than non-pericentromeric regions [13, 28]. Higher methylation levels in non-genic and pericentromeric regions potentially provide a greater amount of base materials for solar-UV induced C→T transition at 5′-Py-$_{m}$CG-3′ context, which is also supported

by our findings of higher frequencies of motifs related to solar-UV signature from methylated regions than unmethylated regions. DNA methylation is highly enriched within transposable elements and repetitive sequences [12, 13, 28]. Thus, this interesting connection between DNA methylation and solar UV-induced mutation propels us to ask a critical question: Is the frequent transition of methylated C to T actually a cost that genomes have to pay for having transposons and repetitive sequences methylated?

Compared with chromosome arms, pericentromeric regions are highly enriched with repetitive sequences and transposable elements, and generally have higher methylation levels, lower gene density, and lower recombination rates [13, 28, 30, 33-36]. In this study, we observed associations between [AT]-difference and methylation level, transposable element, and recombination rate. A previous study illustrated that DNA transposon activity is associated with an increased number of mutations in the sequences close to the transposon [68]. This suggests that enriched transposable elements at pericentromeric regions may contribute to the increased accumulation of mutations within these regions. In sexual organisms, non-recombining regions of genome were found to be subjected to Muller's ratchet [69-72]; and regions with active recombination are more efficient in purging of the deleterious mutations [39]. This may also partially explain the findings of enriched mutations related to solar-UV signature and enlarged [AT]-difference in pericentromeric regions.

Solar UV primarily induces C→T base transition at 5′-Py_mCG-3′ sequence context [20, 40, 73] and CG methylation can enhance solar-UV induced mutation at 5′-Py_mCG-3′ sites [25]. However, a few questions still need to be addressed to understand the increased rate of mutations related to solar-UV signature in domesticated accessions. The first question is how DNA methylation varies across populations as variation in DNA methylation level may lead to the observed difference in rate of mutations related to solar-UV signature between domesticated and

wild groups. A recent study on 51 diverse maize inbred lines identified 172 maize-teosinte differentially methylated regions (DMRs), which are biased toward more examples of higher methylation levels in teosinte than maize [74]. Because those DMRs only represent a very small portion of genome and the majority of methylated regions are conserved within maize, the identified DMRs should not be a major contributor to the observed difference in rate of mutations related to solar-UV signature between two groups. The other question is how UV could induce germ-line mutations as germ-line cells are generally shielded from direct solar radiation. The damaging effects of solar UV are often limited to the epidermis cells due to low UV-B penetration into plant tissues through flavonoid layer [75, 76]. However, some evidences suggest that UV-B may penetrate into meristematic tissues as increased genome instability in plant germline has been observed even with low UV-B radiation [77]. In addition, plant germline cells divide several times during vegetative growth stage and separated into sex-specific lineages only during late flower development [78]. Thus, we suspect that mutations induced by solar UV during vegetative growth in cells of the apical meristem may be inherited into the progeny.

Using a phenotype summarized from millions of SNPs, we identified a set of UV-related genes nearby signals associated with genome divergence. We speculate at some point before domestication, during gametogenesis, spontaneous mutations randomly took place within a UV-related gene. The gene with altered sequence may have mild difference in terms of locating or repairing DNA errors [79]. Therefore, the lineages in which mutations in UV-related genes were segregating began to accumulate systematic difference in DNA repair, which contributed to the genome divergence patterns captured by base composition. In the mutation accumulation experiments, once a *Escherichia coli* lineage acquired 1bp insertion in *mutT* gene at 26,500th generation, the later generations from this lineage began to show greatly elevated mutation rates and bias toward substitution type from A to C than the progenies from other lineages [73]. The

recent study that compared the accumulated mutations after 20 generations between wild type and DNA repair deficient mice suggested different patterns in rate and direction between two lineages [80]. Similar phenomenon has been observed for somatic mutations in cancer cell. The substitution type and rate vary for the patients with different variations in DNA repair genes [81]. The varied mutation rate has been reported in natural populations at the genome level [82], the family level [83], and the subpopulation level [6]. These findings suggested the hypothesis that polymorphisms within UV-related genes played a role in different DNA repair efficiency, which in turn affected the mutation rate differently in different lineages.

Initiation of domestication typically involved a set of key genes controlling for domestication syndrome, a set of traits differentiating wild and domesticated accessions. The causal polymorphisms underlying the domestication syndrome are sought to be the direct targets of artificial selection [1, 3, 4]. Although the UV-related genes were detected through a genome phenotype clearly separated between domesticated and wild accessions, we speculate that these genes were probably not the direct targets because these polymorphisms were less likely to lead to visible agronomic traits that human ancestors desired. The observation that wild and domesticated accessions share the same haplotype for *ATR* and *Lig1* suggested that the polymorphisms in these two genes more likely emerge earlier than the onset of domestication. The consequence of changing these UV-related genes probably promoted the occurrence of desired traits, which was subject to the direct selection. The identified UV-related genes indicate almost every step in NER pathway contributes to the overall [AT]-increase (Figure S23), suggesting the complexity of molecular mechanisms.

Molecular experiments need to be carried out to provide evidence supporting the function of these UV-related genes and their connection to the base-composition pattern. Although it is beyond the scope of this study to address the functional difference between wild and

domesticated alleles and the molecular mechanisms affecting the repair efficiency, this study pointed to a new direction for addressing some fundamental questions about the genome itself. We think that mutation repair genes, like *ATR* and *Lig1*, harboring significant changes such as altered gene structure, should be the next priority to study and provide molecular evidences. Induced mutation accumulation experiments with UV as the mutagen and near isogenic lines (NILs) segregating only at regions surrounding mutation repair genes as starting materials will be preferable to demonstrate the connection between UV-induced mutation and base composition change. Sequencing lines that derived from starting materials carrying mutations at UV damaged DNA repair gene regions may also provide additional support.

Conclusions

Base-composition difference between domesticated accessions and wild accessions at the dynamic part of the genome suggests the important role of AT-bias mutation in shaping overall pattern of base-composition variation. Regional variations of base-composition pattern indicate that non-genic SNPs and pericentromeric regions have greater contributions to the observed pattern. This finding together with the discovery of solar UV's potential role in driving the genome divergence establish the connection between DNA methylation and base-composition variation. By focusing on the evolutionary outcome, our genome scans in maize and soybean identified a set of UV damage repair genes. Rapidly improved genomics and epigenomics capacity would further facilitate our efforts to probe potential connections among base-composition, mutation, methylation, DNA repair, and genome evolution.

Author Contributions

JY, XL, and JW designed the study. JW, XL, KKD, MJS, SAJ, and NMS conducted the analyses. JW, XL, and JY wrote the manuscript with inputs from all authors. All authors read and approved the final manuscript.

Figures

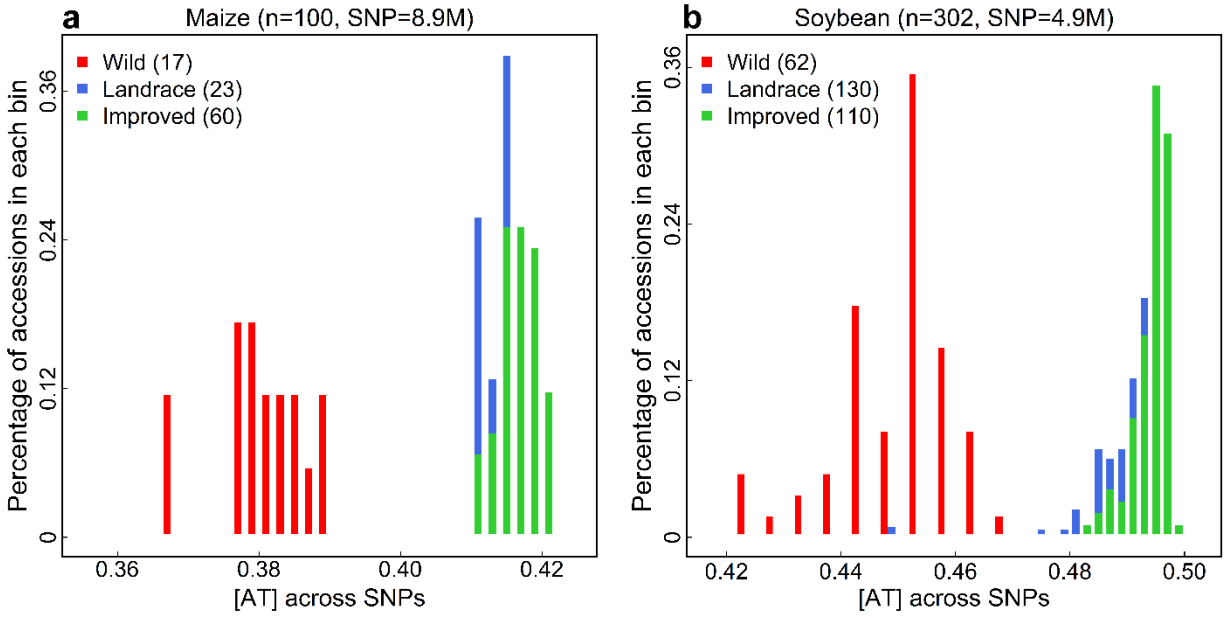


Figure 1. Genome-wide base-composition pattern in maize and soybean. (**a**) The distribution of [AT] among 8.9 million SNPs in 100 maize accessions. (**b**) The distribution of [AT] across 4.9 million SNPs in 302 soybean accessions.

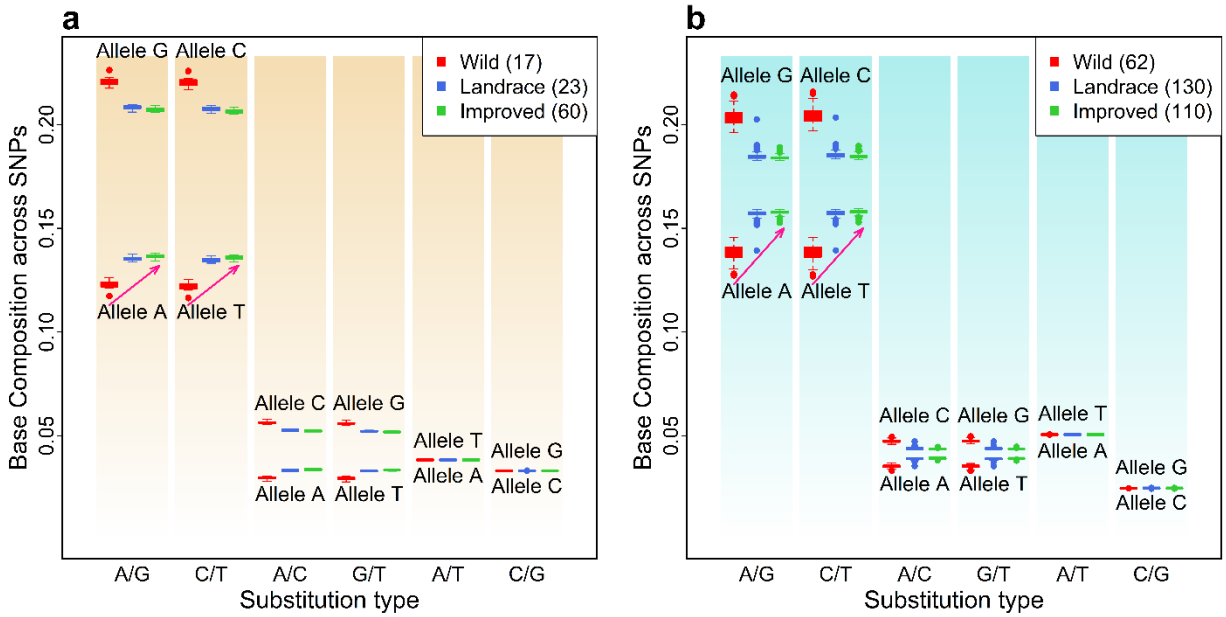


Figure 2. Base-composition distribution at each of the 6 substitution types in maize (**a**) and soybean (**b**). The genome-wide SNPs were classified into 6 substitution types. Base composition was calculated for each accession conditional on each substitution type. The red arrows show the [A] and [T] increase at A/G and C/T substitution types.

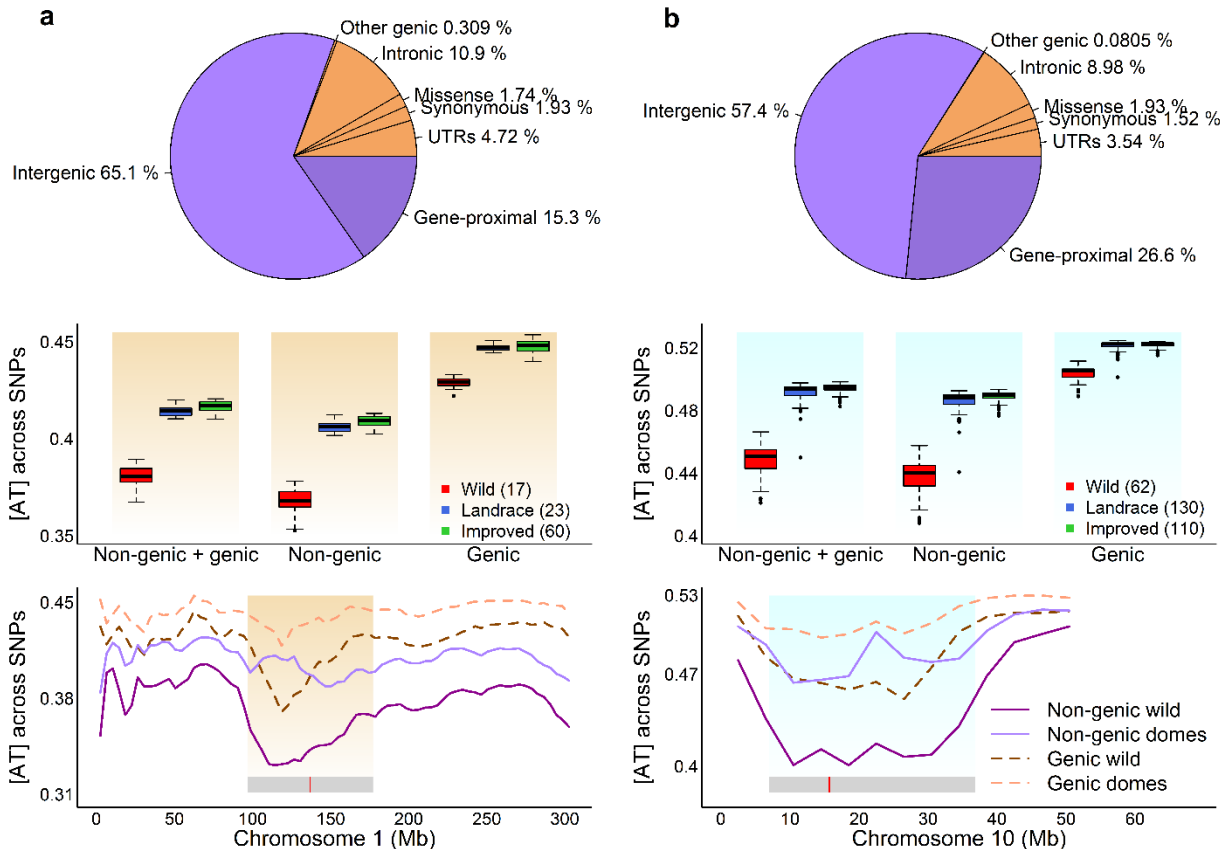


Figure 3. The distribution of base composition calculated with genic and non-genic SNPs in maize (**a**) and soybean (**b**). The upper panel shows the distribution of SNPs across different genomic annotation sets. The middle panel shows the base-composition distribution with genic and non-genic SNPs. The lower panel illustrates the base-composition distribution across 5Mb segments with genic and non-genic SNPs. To simplify the plot in the lower panel, landraces and improved cultivars are combined to be domesticated group to compare with wild group. For each accession, base-composition was calculated using a moving average approach with a 5-Mb window size and a 4-Mb step size. Each point in the plot represents the mean [AT] of the specified group across a 5-Mb window. The gray bar in the bottom indicates the position of pericentromeric region, and the red bar within gray bar shows the position of centromeric region.

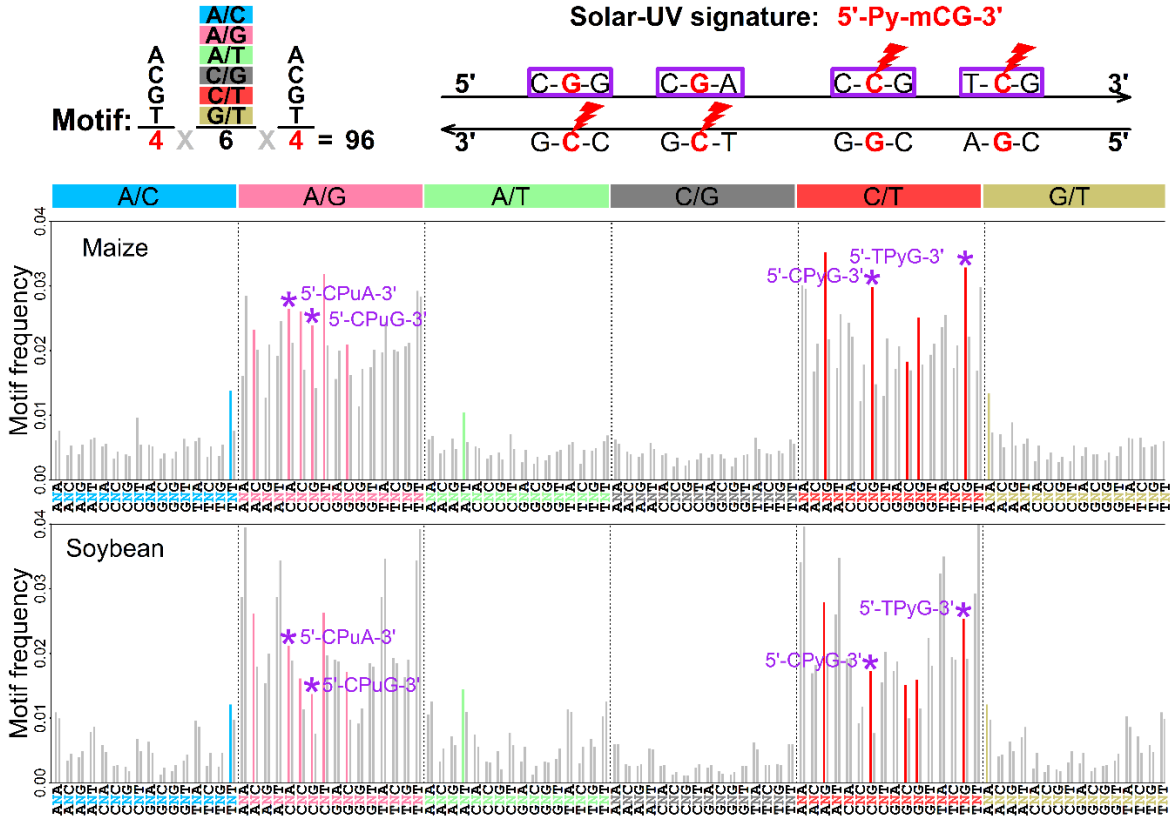


Figure 4. Motif enrichment analysis in maize and soybean. The upper panel illustrates the composition of tri-nucleotide motifs and the induction of motifs related to solar UV-signature on double strand DNA. Each tri-nucleotide motif is formed by incorporating reference base pairs immediately upstream and downstream to middle SNP site. Ninety-six motifs are divided to 6 classes based on the substitution types of the SNP. The lightning sign shows the mutation site, and the purple rectangle highlights motifs related to solar-UV signature. The middle and lower panel show the frequency of motif in maize and soybean, respectively. For each motif, the left bar is the overall frequency around SNP sites, while the right bar is the overall frequency of the same motif around random sites (An empirical 95th percentile drawn from 100 random sample scenarios). The colored bar indicates the common motif between maize and soybean with frequency greater than 1/96 and the frequency of motif at SNP sites is higher than that at random sites. The bar with a star on top highlights the motif related to solar-UV signature.

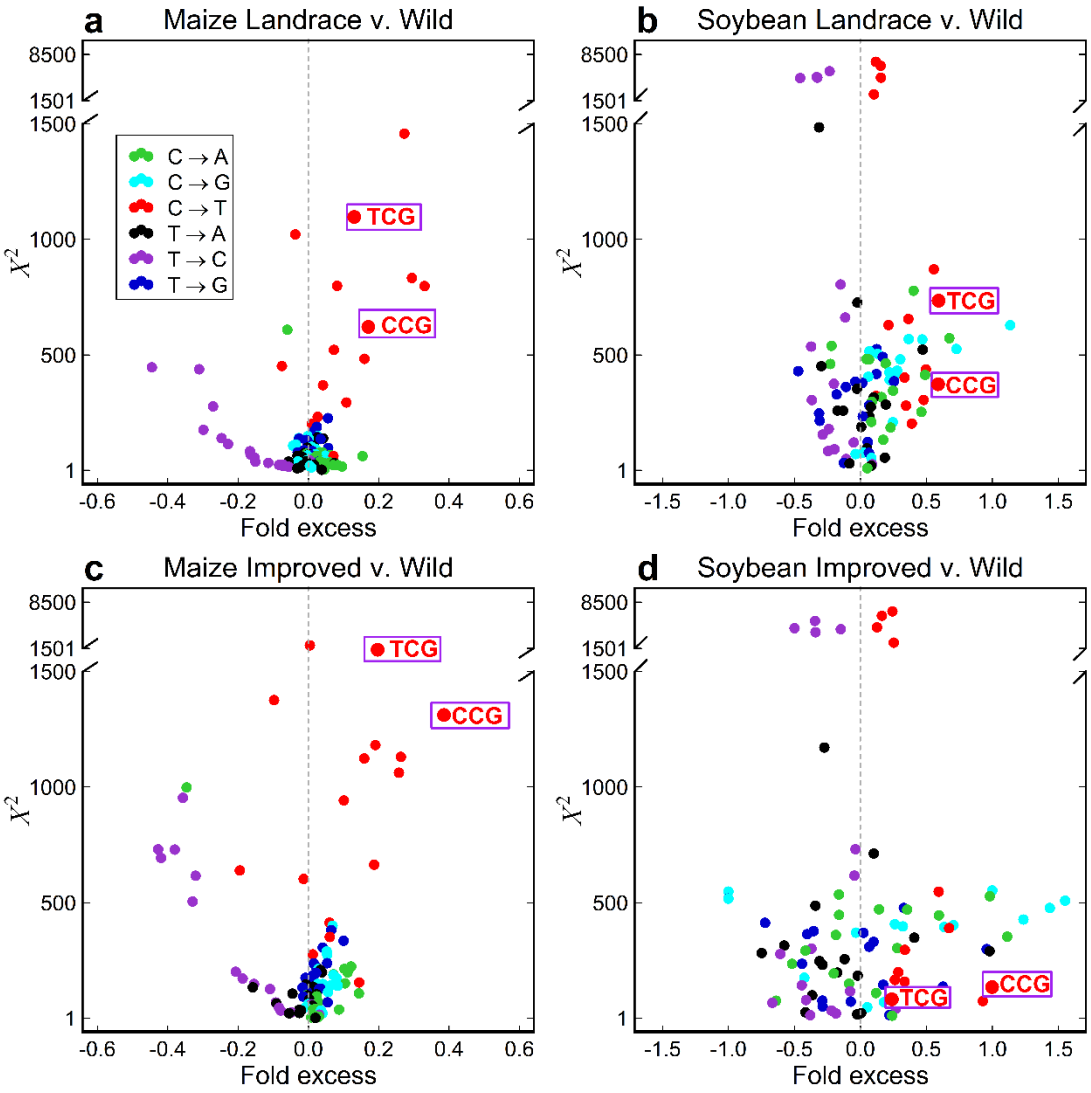


Figure 5. Enrichment test of mutations related to solar-UV signature with population-private SNPs. **a**, **b** Compare the mutation frequency between landraces and wild accessions in maize and soybean respectively, and the x coordinate of each point indicates the fold frequency difference $(f_{PL}(m) - f_{PW}(m))/f_{PW}(m)$. **c**, **d** Compare the mutation frequency between improved cultivars and wild accessions in maize and soybean respectively, and the x coordinate of each point indicates the fold frequency difference $(f_{PI}(m) - f_{PW}(m))/f_{PW}(m)$. The y coordinate indicates the Pearson's χ^2 value that measures the significance of the difference between $f_m(P_1)$ and $f_m(P_2)$. Outlier points are labeled with the ancestral state of the mutant nucleotide flanked by two neighboring bases, and the color of the points indicate the ancestral and derived alleles of the mutant site. The purple rectangle highlights the mutations related to solar-UV signature. Here TCG on the plot represents mutation 5′-TCG-3′→5′-TTG-3′ and its reverse complement 5′-CGA-3′→5′-CAA-3′, CCG represents mutation 5′-CCG-3′→5′-CTG-3′ and its reverse complement 5′-CGG-3′→5′-CAG-3′, and similarly for all the other dots on the plot.

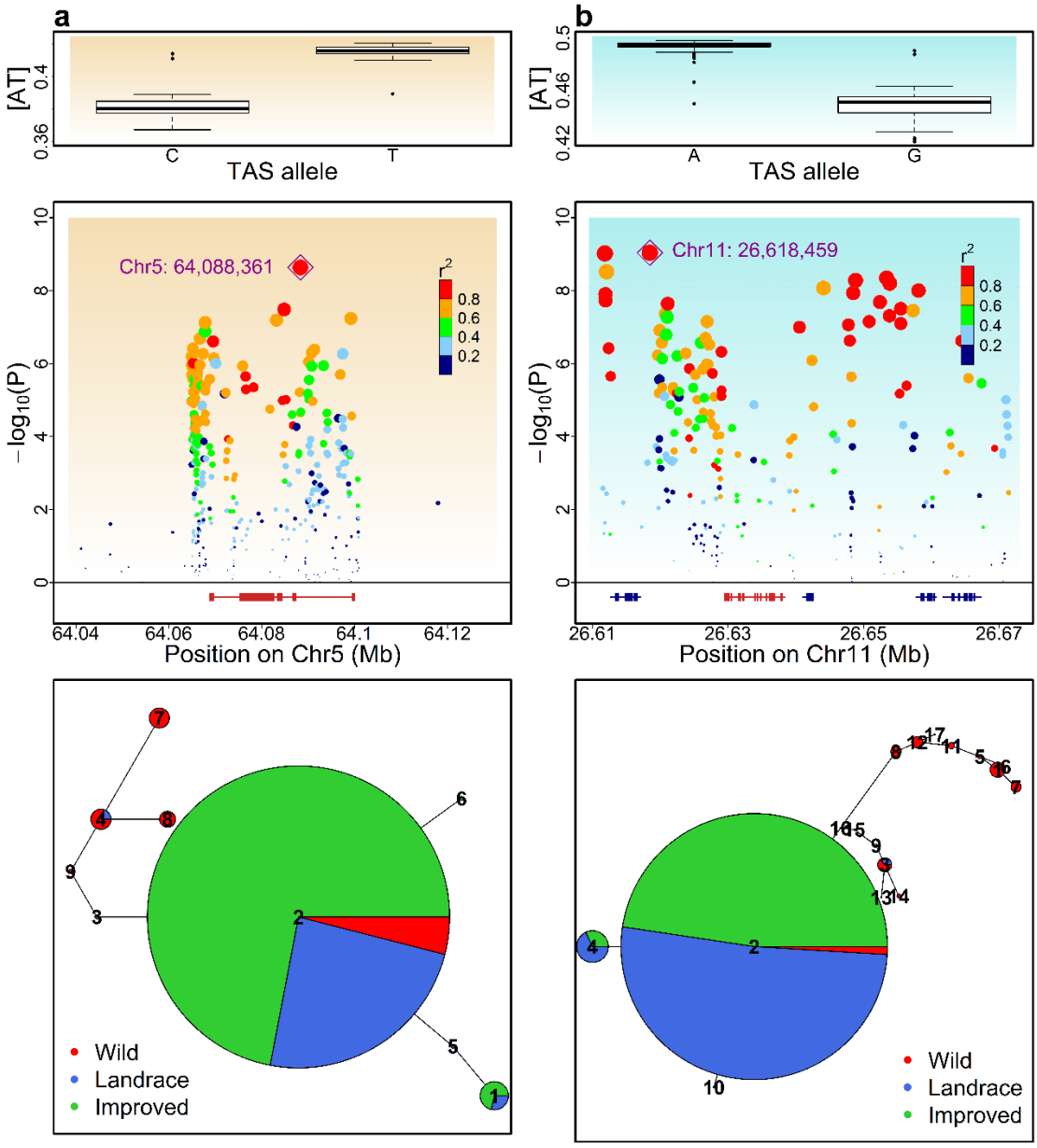


Figure 6. UV related DNA repair genes implicated by trait-associated SNPs (TASs) and haplotype demographic distributions. (**a**) *ATR* in maize is tagged by a TAS (PZE0561610418) on chromosome 5. (**b** *DNA ligase1* (*Lig1*) in soybean is tagged by a TAS (rs1126618459) on chromosome 11. The upper panel shows the boxplot of base composition between accessions carrying different alleles at the TASs. The middle panel shows the regional Manhattan plot around *ATR* and *Lig1* locus (*ATR* and *Lig1* are show in red, others in blue). Dot size is proportional to the magnitude of significance for the SNP's association with [AT] variation. Dot color indicates its LD with the TAS. The lower panel shows the haplotype networks inferred from 8 SNPs within *ATR* gene and 16 SNPs within *Lig1* gene, respectively. Each circle represents one haplotype. Size of the circle is proportional to the number of accessions possessed the haplotype. Size of each colored slice within a circle is proportional to the number of accessions possessed the haplotype from the corresponding group.

Supplementary Information

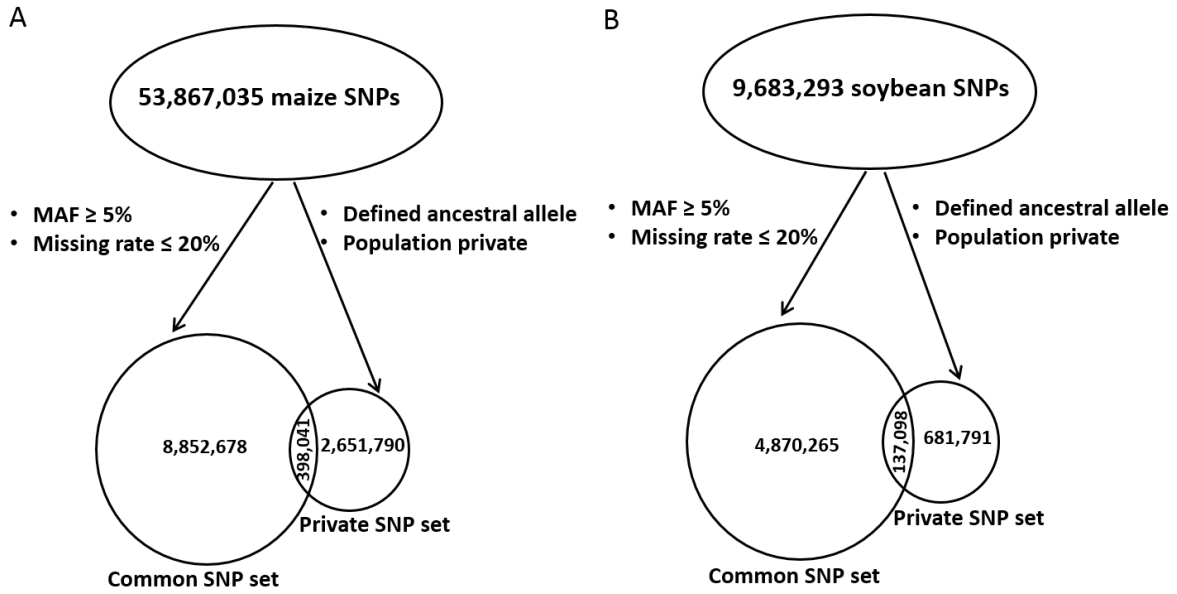


Figure S1. Diagram of SNP filtering process in (**A**) maize and (**B**) soybean. The common SNP set is filtered with a minor allele frequency (MAF) threshold of 5% and a missing rate threshold of 20%. The private SNP set is obtained by identifying SNPs that have defined ancestral allele information and meet the population-private SNP definition.

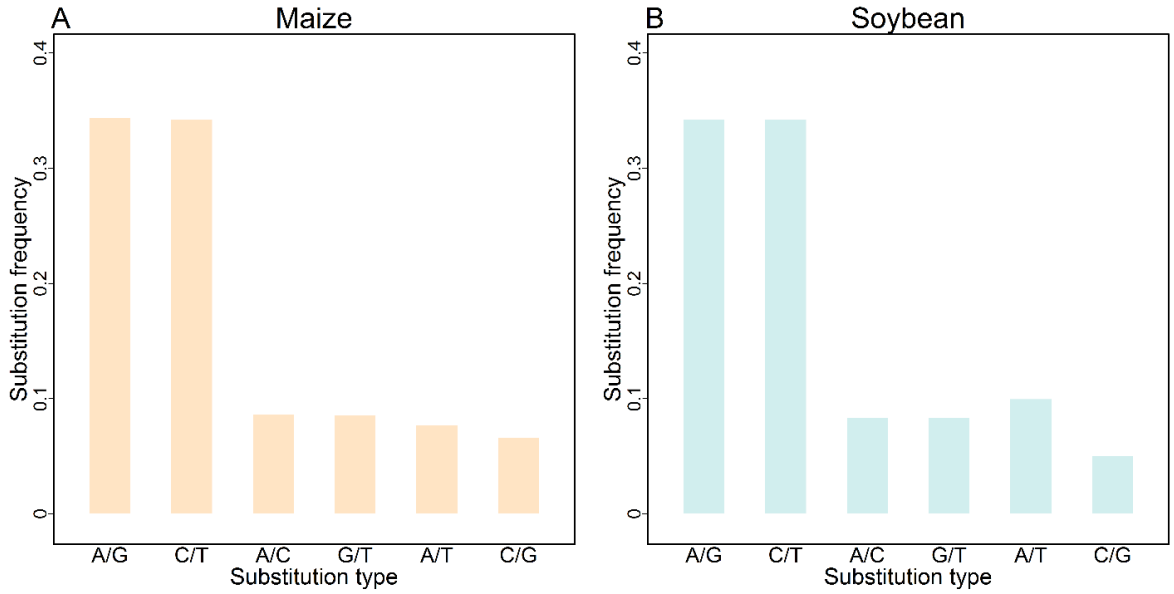


Figure S2. Frequency of substitution types among detected SNPs in (**A**) maize and (**B**) soybean. The genome-wide SNPs were classified into 6 substitution types. Then frequency for each of the 6 substitution types was calculated.

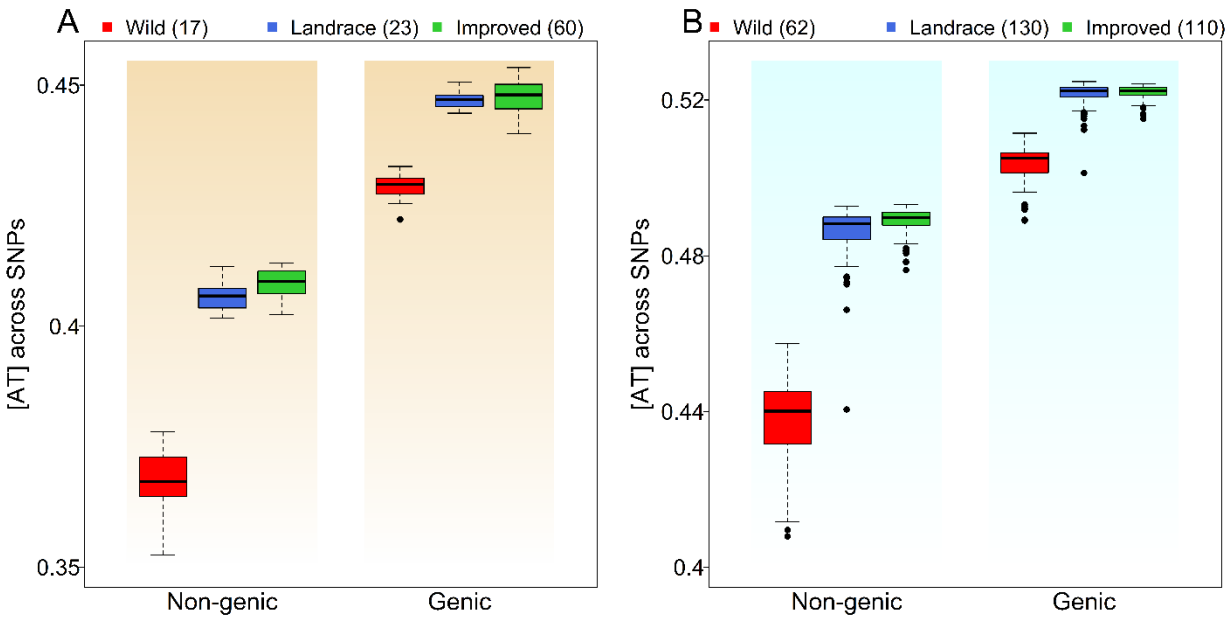


Figure S3. Base-composition distribution for randomly sampled non-genic SNPs and genic SNPs. (**A**) The distribution of [AT] calculated with 1.56 million non-genic SNPs and 1.56 million genic SNPs in maize. (**B**) The distribution of [AT] calculated with 0.7 million non-genic SNPs and 0.7 million genic SNPs in soybean. An equal amount of SNPs were randomly sampled from genic and non-genic SNP sets. The average [AT] values over 100 iterations were plotted.

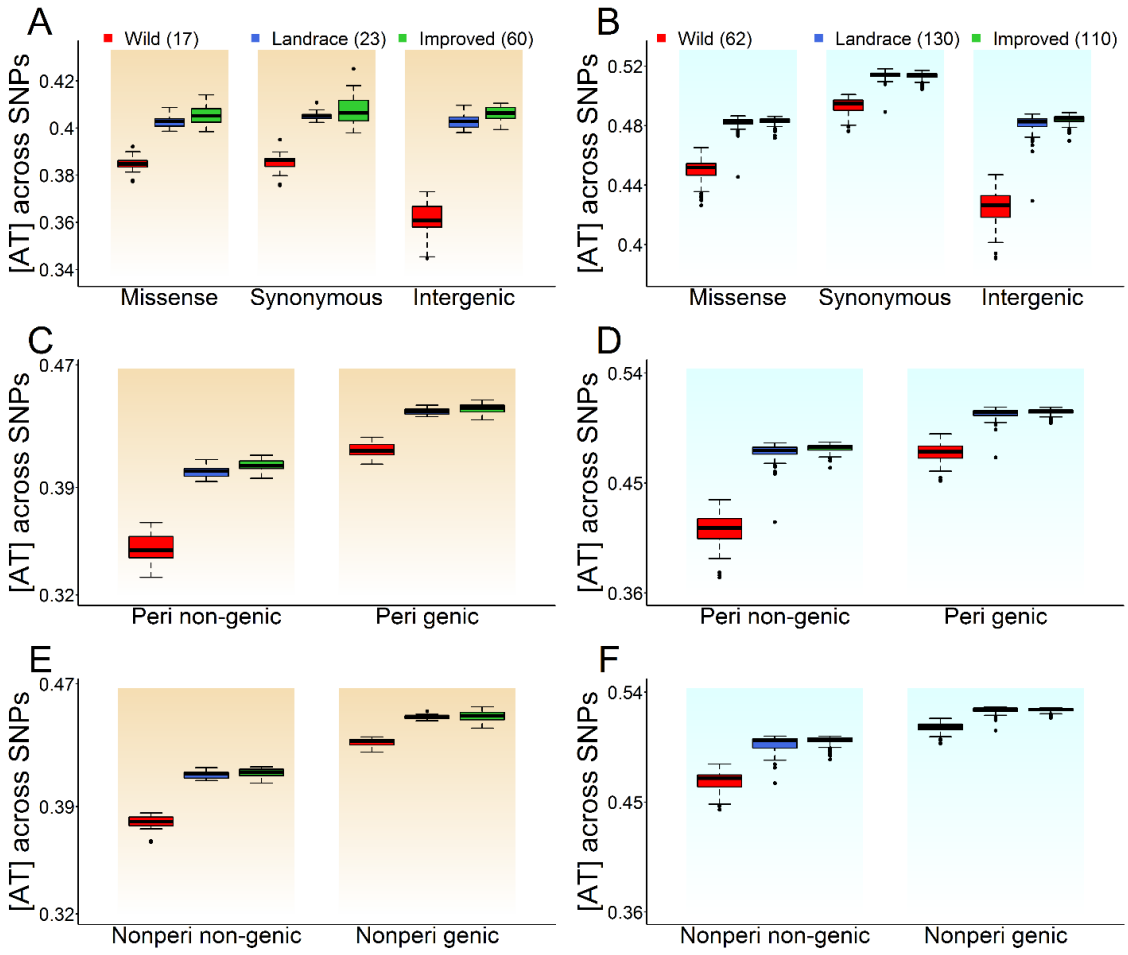


Figure S4. Comparison of base composition distribution between different regions of the genome. (**A**) Comparison between missense, synonymous, and intergenic SNP sets in maize. (**B**) Comparison between missense, synonymous, and intergenic SNP sets in soybean. (**C**) Comparison between non-genic and genic regions conditional on pericentromeric regions in maize. (**D**) Comparison between non-genic and genic regions conditional on pericentromeric regions in soybean. (**E**) Comparison between non-genic and genic regions conditional on chromosome arms in maize. (**F**) Comparison between non-genic and genic regions conditional on chromosome arms in soybean.

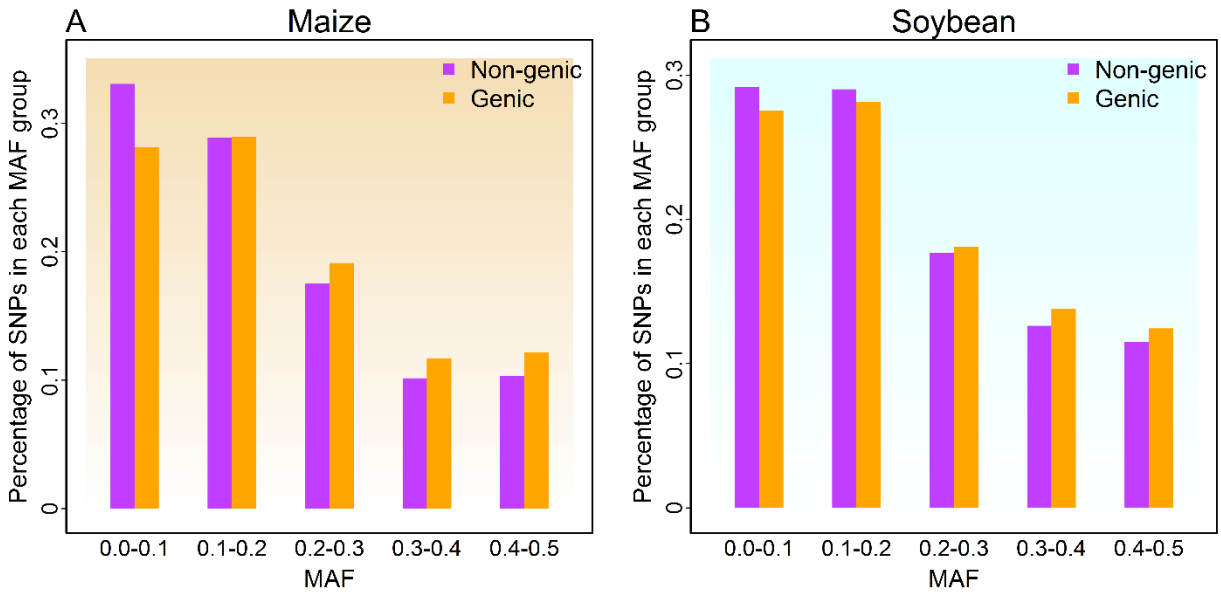


Figure S5. The distribution of minor allele frequency (MAF) for genome-wide genic and non-genic SNPs in (**A**) maize and (**B**) soybean. The percentage of SNPs fallen in each MAF bin was calculated for genic and non-genic SNPs, respectively.

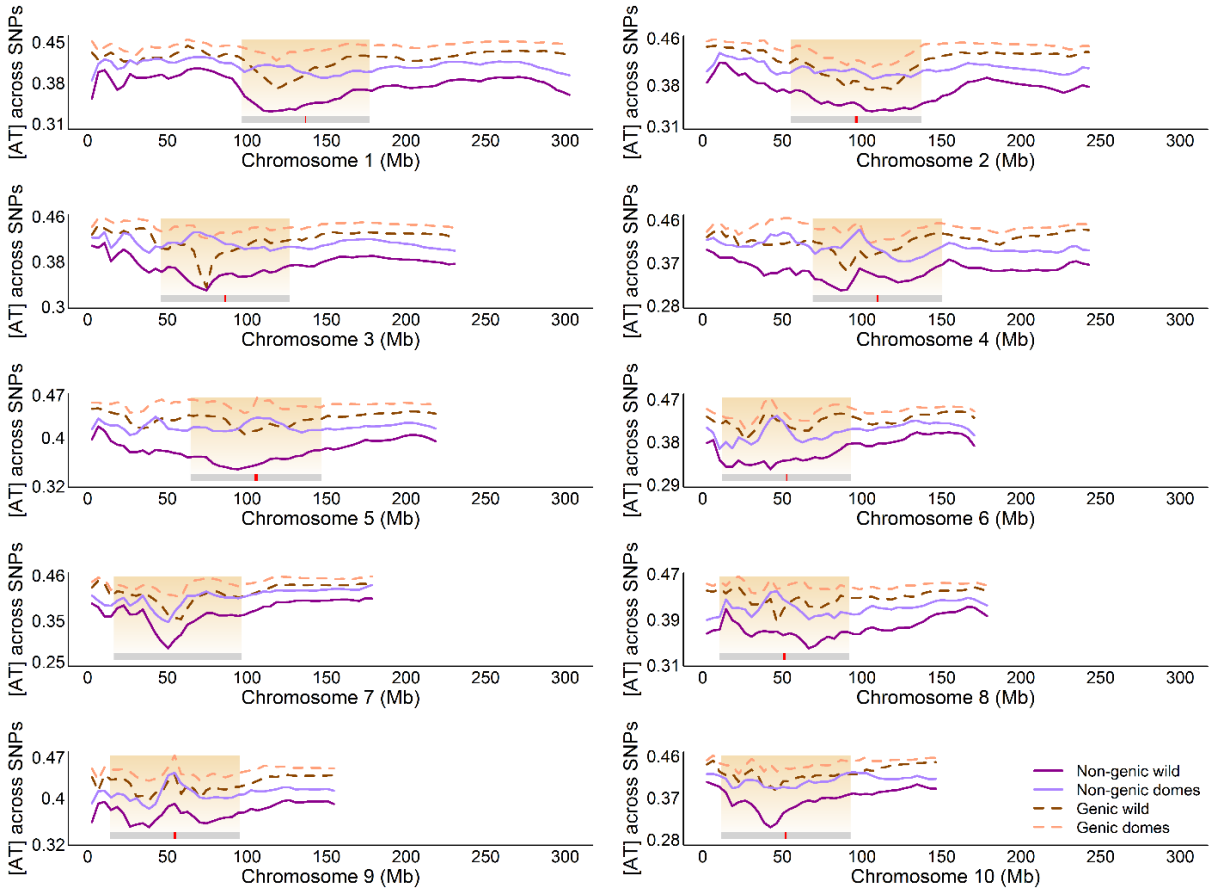


Figure S6. The distribution of base composition calculated with genic and non-genic SNPs across 5Mb segments for each of 10 maize chromosomes. Landraces and improved cultivars are combined to be domesticated group to compare with wild group. For each accession, base-composition was calculated using a moving average approach with a 5-Mb window size and a 4-Mb step size. Each point in the plot represents the mean [AT] of the specified group across a 5-Mb window. The gray bar in the bottom indicates the position of pericentromeric region, and the red bar within gray bar shows the position of centromeric region.

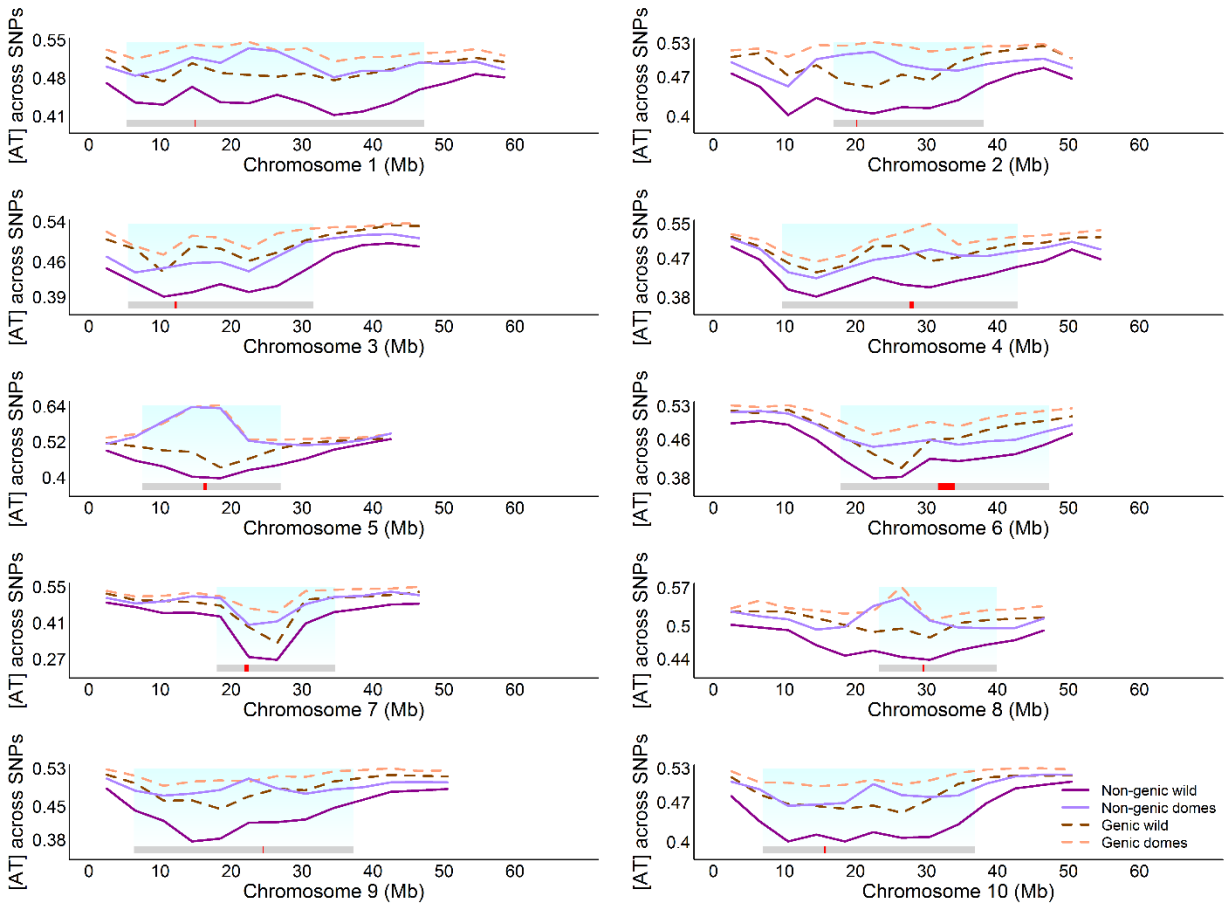


Figure S7. The distribution of base composition calculated with genic and non-genic SNPs across 5Mb segments for soybean chromosome 1-10. Landraces and improved cultivars are combined to be domesticated group to compare with wild group. For each accession, base-composition was calculated using a moving average approach with a 5-Mb window size and a 4-Mb step size. Each point in the plot represents the mean [AT] of the specified group across a 5-Mb window. The gray bar in the bottom indicates the position of pericentromeric region, and the red bar within gray bar shows the position of centromeric region. The high [AT] at pericentromeric region (10-20 Mb) of chromosome 5 for domesticated accessions (and the resulting larger [AT]-difference compared with wild accessions) may be due to an extensive selective sweep region detected in this region (Nat Biotechnology 2015, 33:408-414).

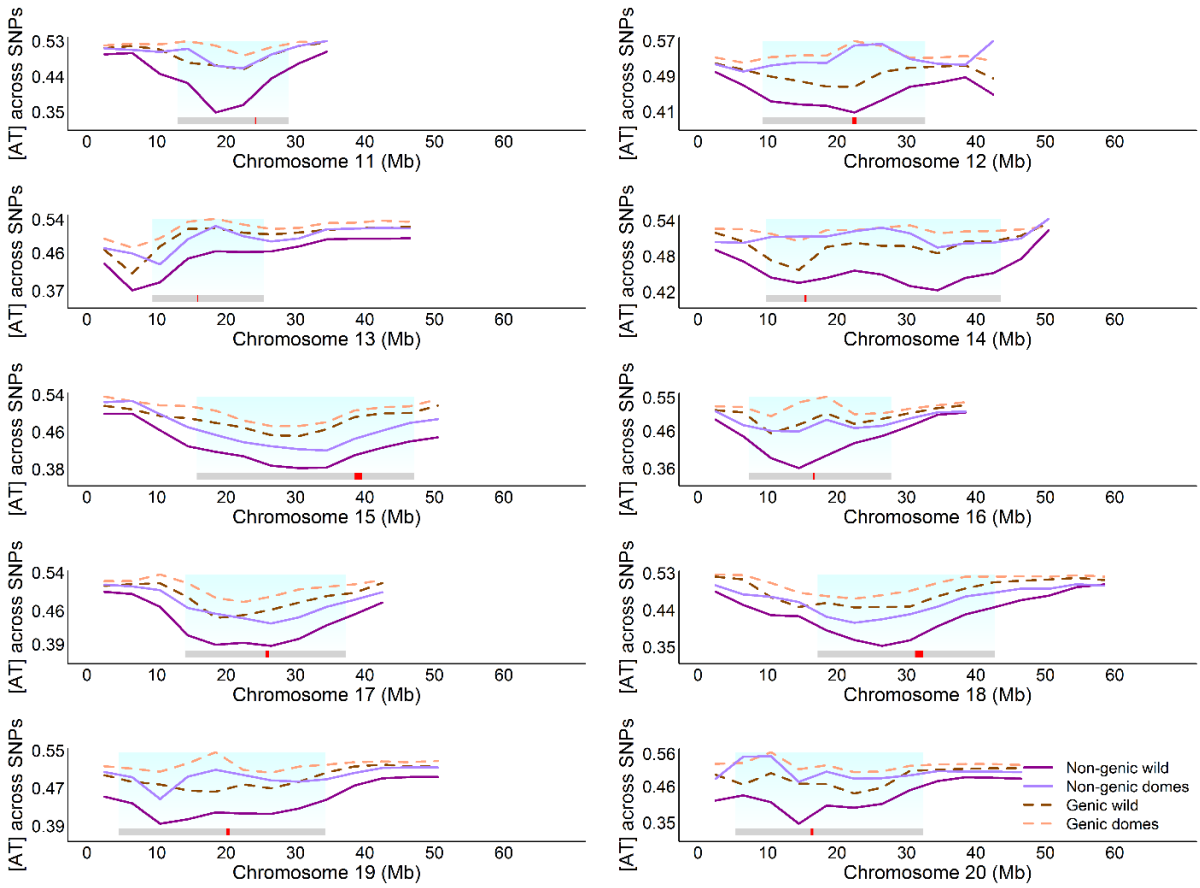


Figure S8. The distribution of base composition calculated with genic and non-genic SNPs across 5Mb segments for soybean chromosome 11-20. Landraces and improved cultivars are combined to be domesticated group to compare with wild group. For each accession, base-composition was calculated using a moving average approach with a 5-Mb window size and a 4-Mb step size. Each point in the plot represents the mean [AT] of the specified group across a 5-Mb window. The gray bar in the bottom indicates the position of pericentromeric region, and the red bar within gray bar shows the position of centromeric region.

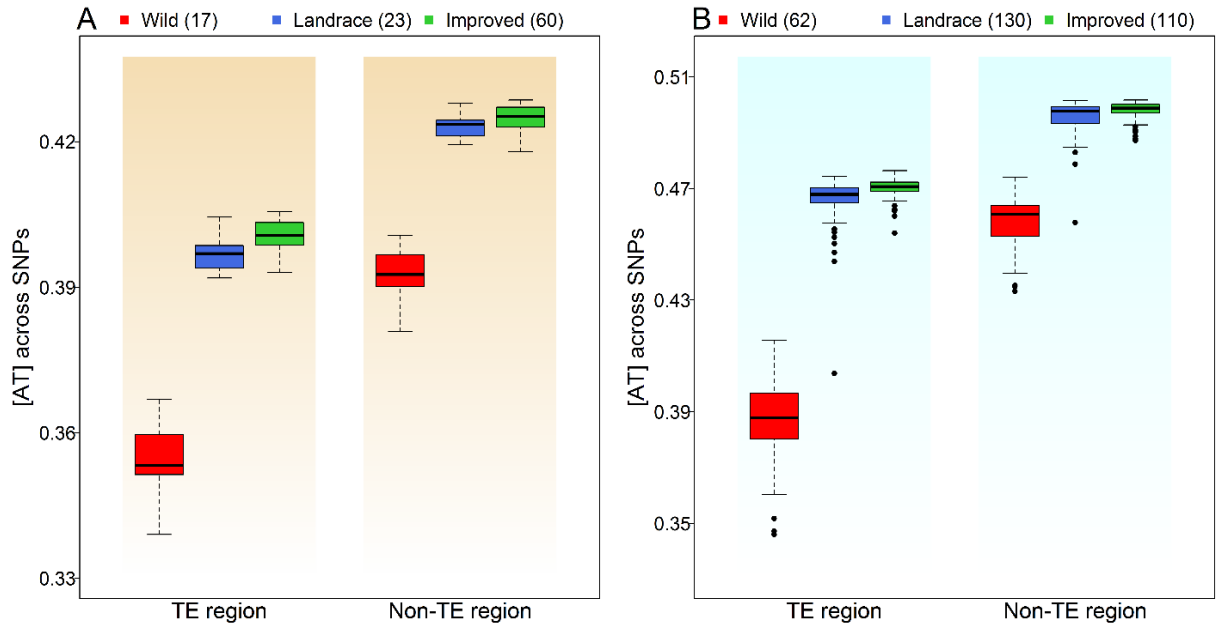


Figure S9. Base-composition distribution at transposable element (TE) and non-TE regions in maize (**A**) and soybean (**B**). The genome-wide SNPs were classified into TE and non-TE regions. And then [AT] were calculated from SNPs of TE and non-TE regions separately.

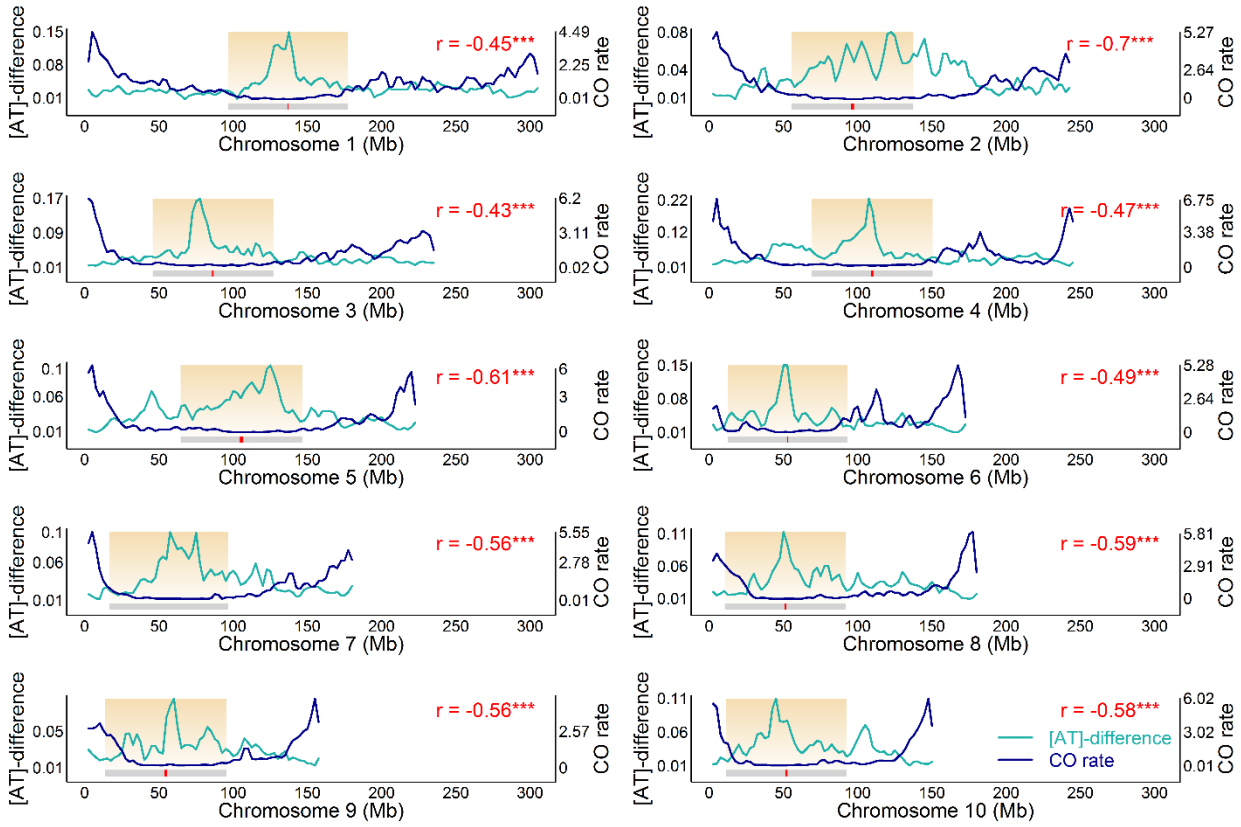


Figure S10. The distribution of base composition between domesticated and wild accessions and Crossover (CO) rate for each of 10 maize chromosomes. Both [AT]-difference CO rate are calculated using a 5-Mb sliding window. *r,* Pearson correlation coefficient between [AT]-difference and CO rate for each of the chromosomes; *, *P*-value ≤ 0.05; **, *P*-value ≤ 0.01, ***, *P*-value ≤ 0.001.

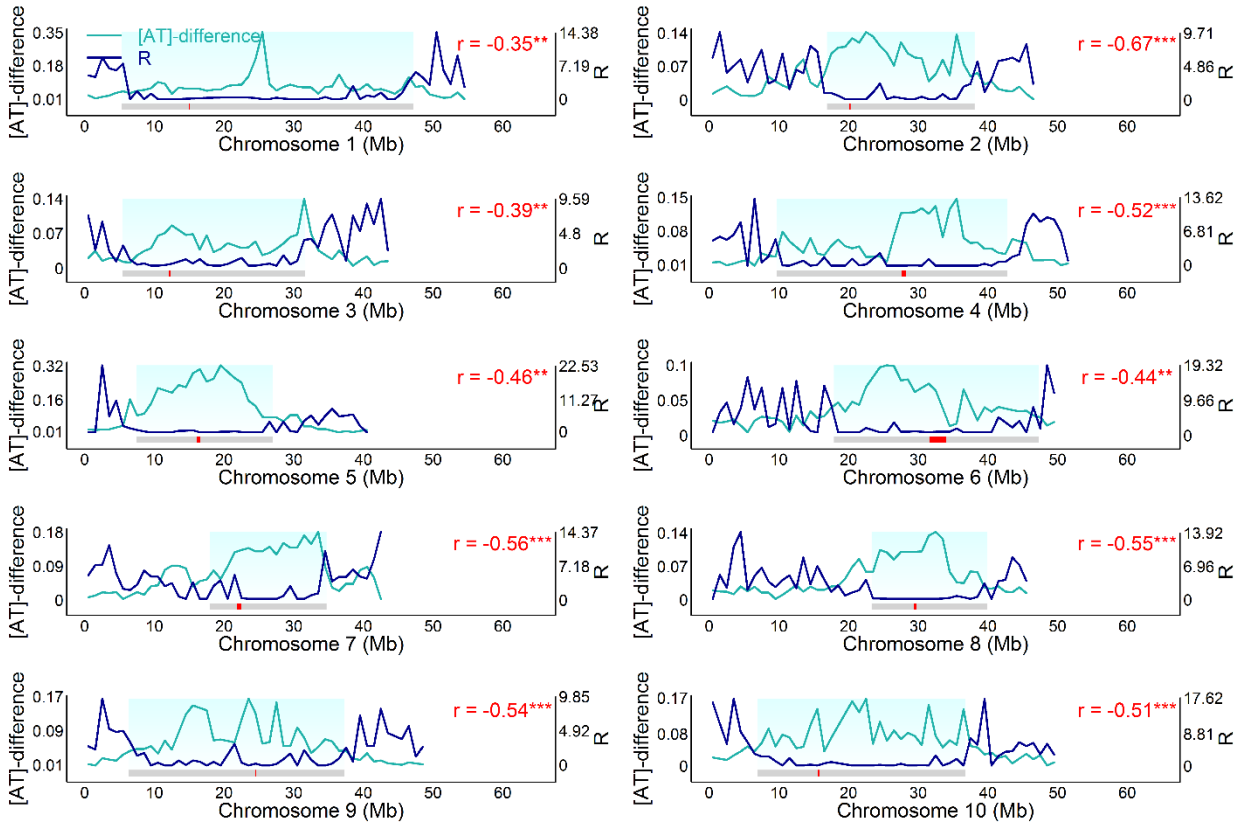


Figure S11. The distribution of base composition between domesticated and wild accessions and genetic recombination rate (R) for each of soybean chromosome 1-10. Both [AT]-difference and recombination rate are calculated using a 1-Mb window. *r,* Pearson correlation coefficient between [AT]-difference and recombination rate for each of the chromosomes; *, *P*-value ≤ 0.05; **, *P*-value ≤ 0.01, ***, *P*-value ≤ 0.001.

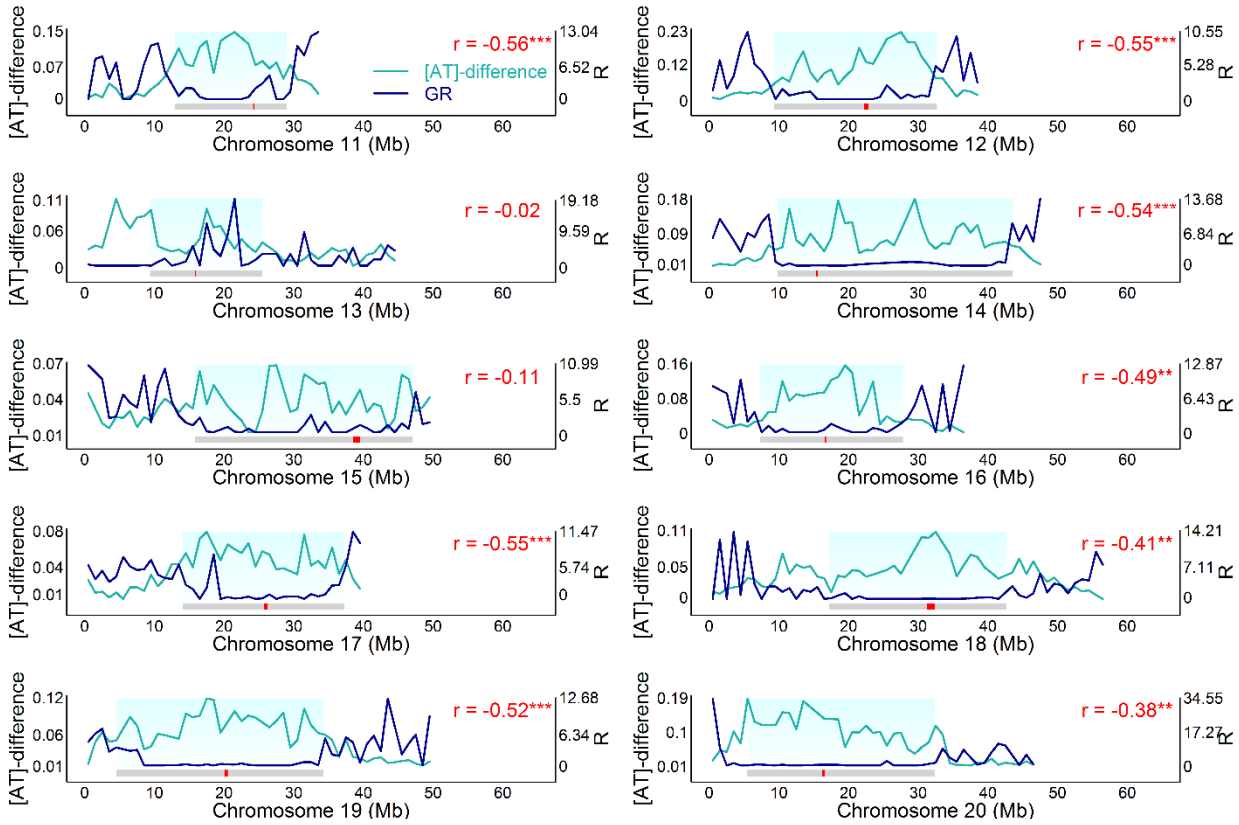


Figure S12. The distribution of base composition between domesticated and wild accessions and genetic recombination rate (R) for each of soybean chromosome 11-20. Both [AT]-difference and recombination rate are calculated using a 1-Mb window. *r,* Pearson correlation coefficient between [AT]-difference and recombination rate for each of the chromosomes; *, *P*-value ≤ 0.05; **, *P*-value ≤ 0.01, ***, *P*-value ≤ 0.001.

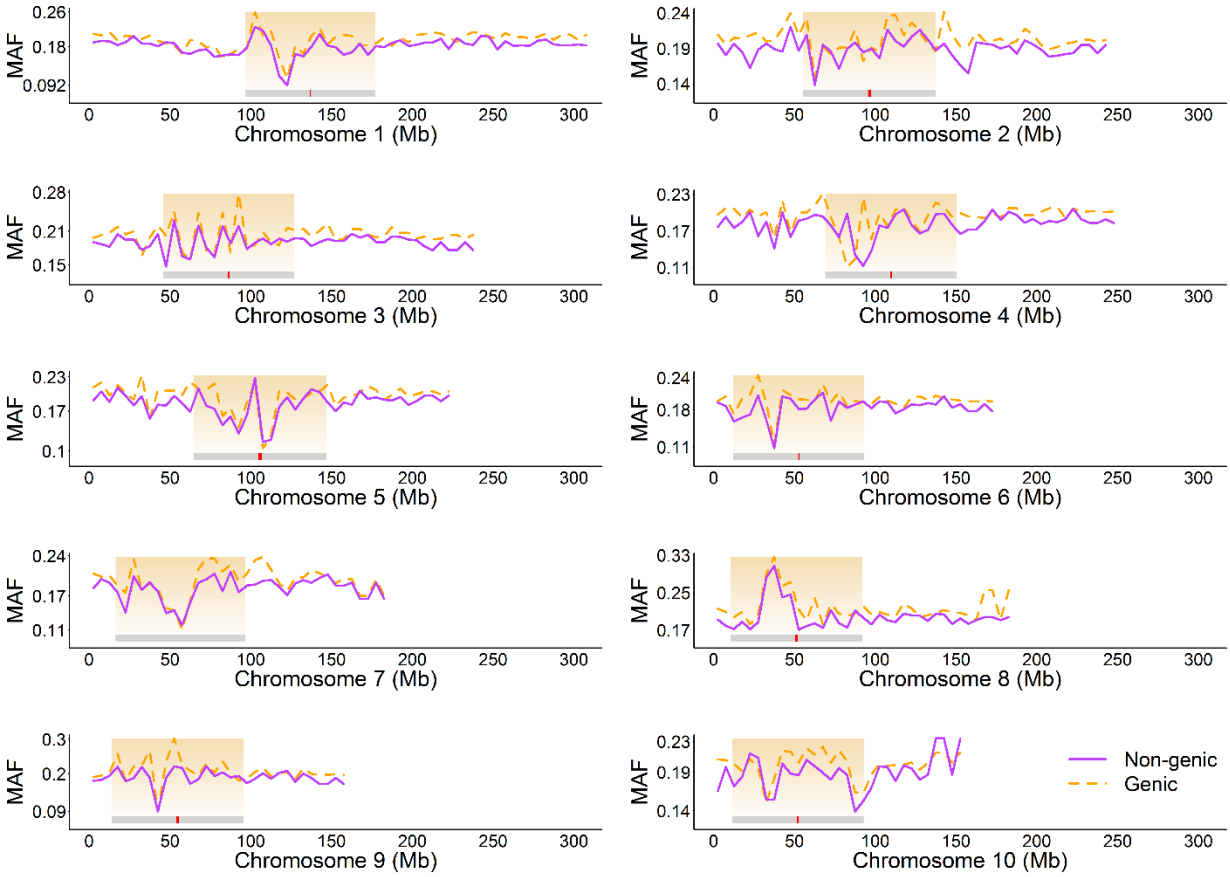


Figure S13. The distribution of minor allele frequency (MAF) calculated with genic and non-genic SNPs across 5Mb segments for each of 10 maize chromosomes. The mean MAF of SNPs was calculated using a moving average approach with a 5-Mb window size and a 4-Mb step size. The gray bar in the bottom indicates the position of pericentromeric region, and the red bar within gray bar shows the position of centromeric region.

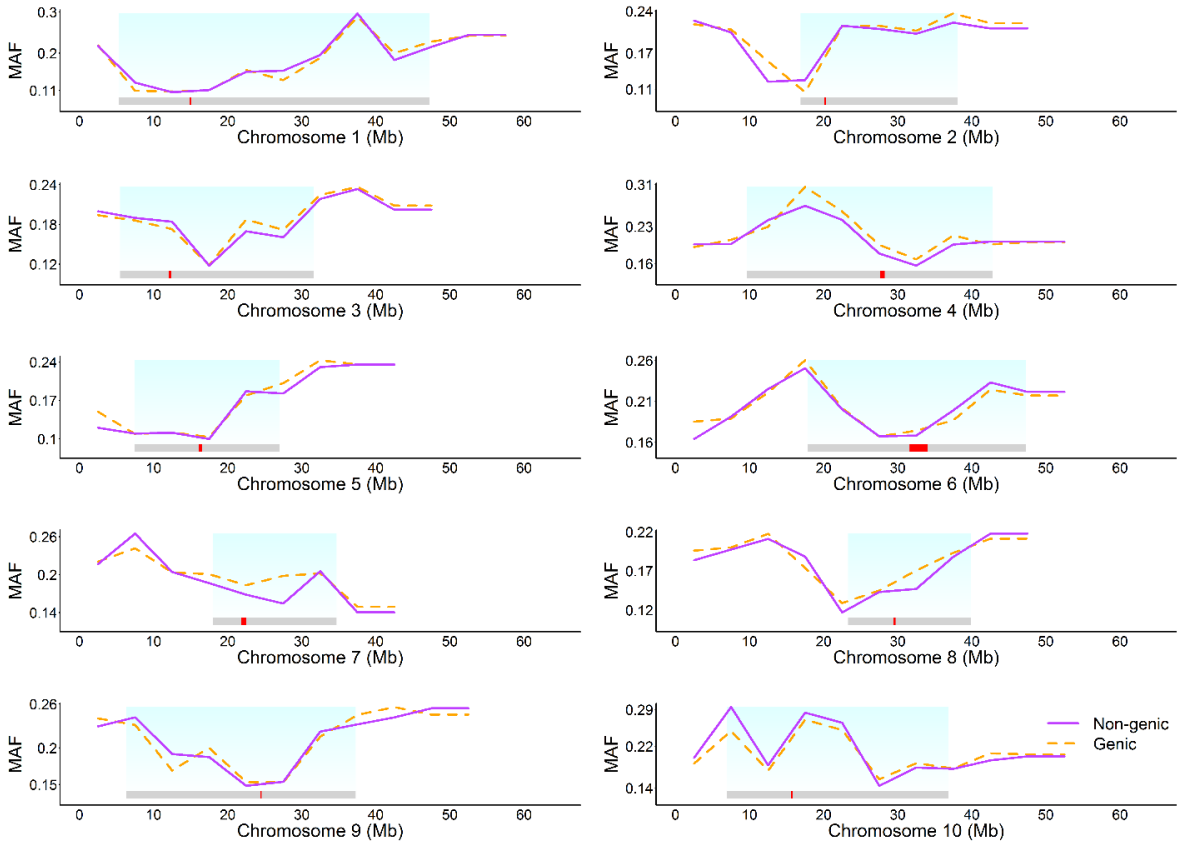


Figure S14. The distribution of minor allele frequency (MAF) calculated with genic and non-genic SNPs across 5Mb segments for soybean chromosome 1-10. The mean MAF of SNPs was calculated using a moving average approach with a 5-Mb window size and a 4-Mb step size. The gray bar in the bottom indicates the position of pericentromeric region, and the red bar within gray bar shows the position of centromeric region.

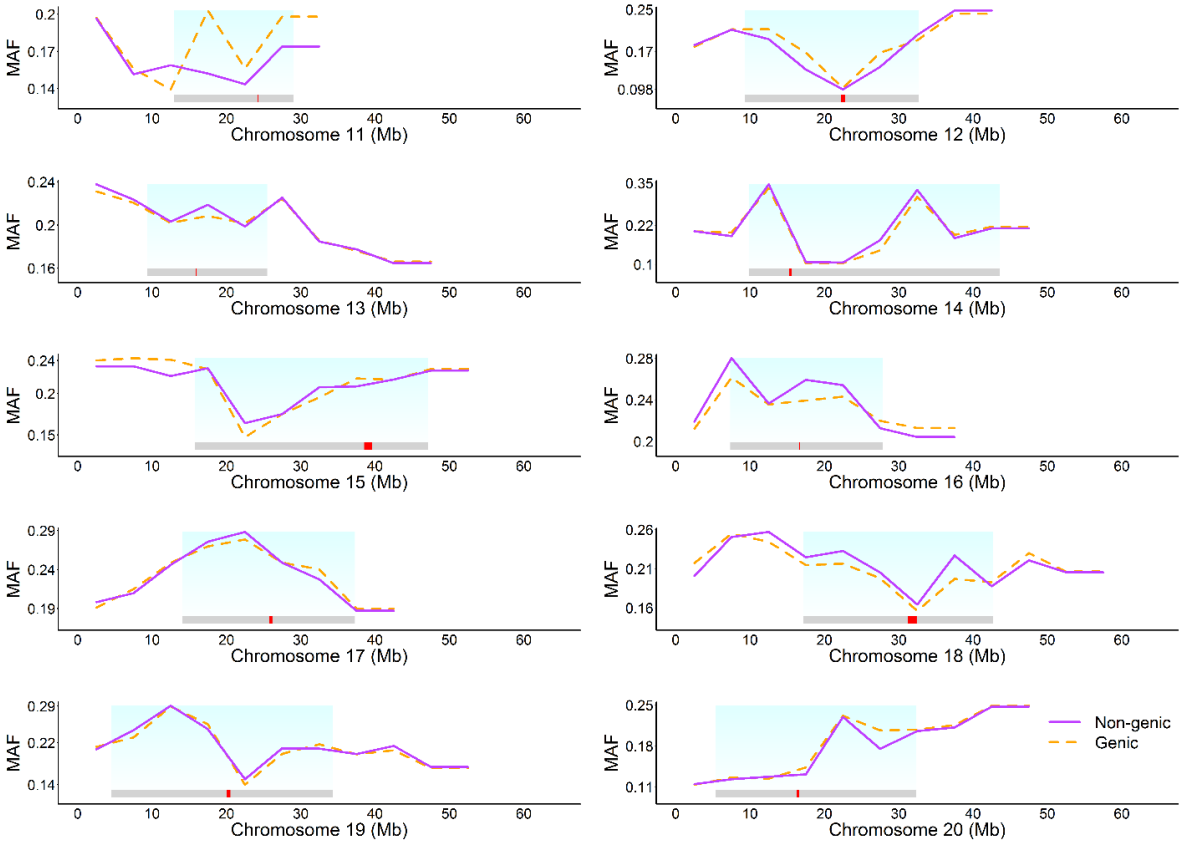


Figure S15. The distribution of minor allele frequency (MAF) calculated with genic and non-genic SNPs across 5Mb segments for soybean chromosome 11-20. The mean MAF of SNPs was calculated using a moving average approach with a 5-Mb window size and a 4-Mb step size. The gray bar in the bottom indicates the position of pericentromeric region, and the red bar within gray bar shows the position of centromeric region.

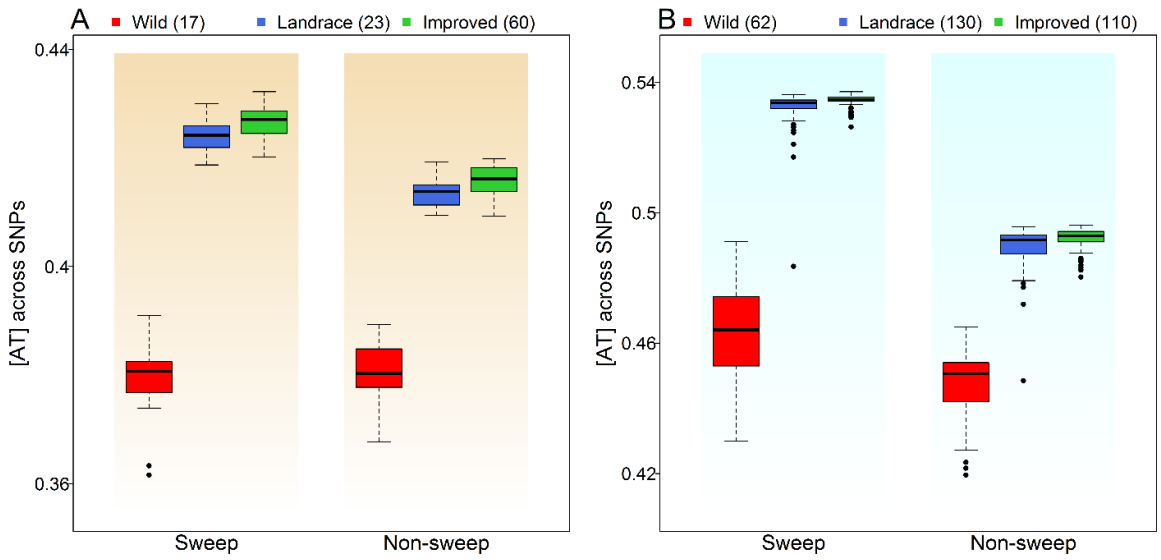


Figure S16. Base-composition distribution at selective sweep and non-selective-sweep regions in maize (**A**) and soybean (**B**). The genome-wide SNPs were classified into selective sweep and non-selective-sweep regions. And then [AT] were calculated from SNPs of selective sweep and non-selective-sweep regions separately.

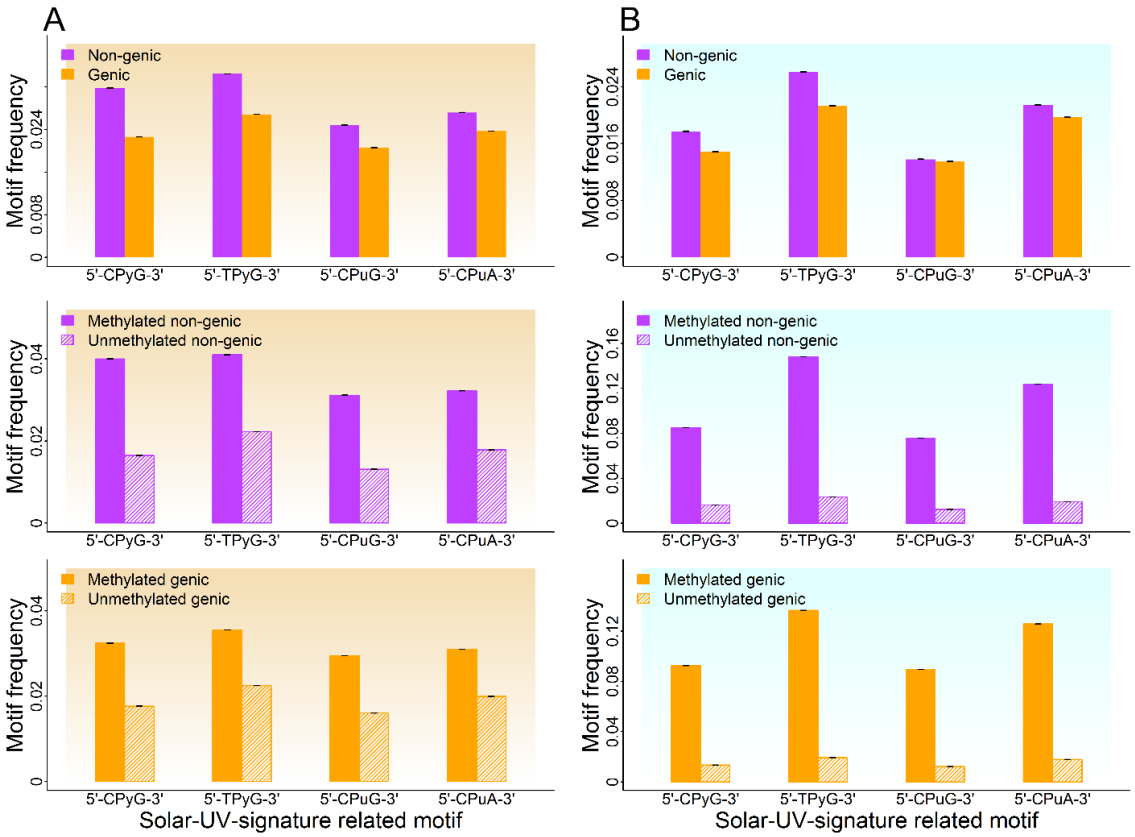


Figure S17. Frequencies of motifs related to solar-UV signature among genic and non-genic SNPs conditional on methylated and unmethylated regions in (**A**) maize and (**B**) soybean. The top panel shows the frequencies of motifs with all genic and non-genic SNPs. The middle panel shows the frequencies of motifs with non-genic SNPs conditional on methylated and unmethylated regions. The bottom panel shows the frequencies of motifs with genic SNPs conditional on methylated and unmethylated regions. Each bar represents the the average frequency of a specific motif over 100 maize accessions in (**A**) and 302 soybean accessions in (**B**). The black bar on the top illustrates the standardard error.

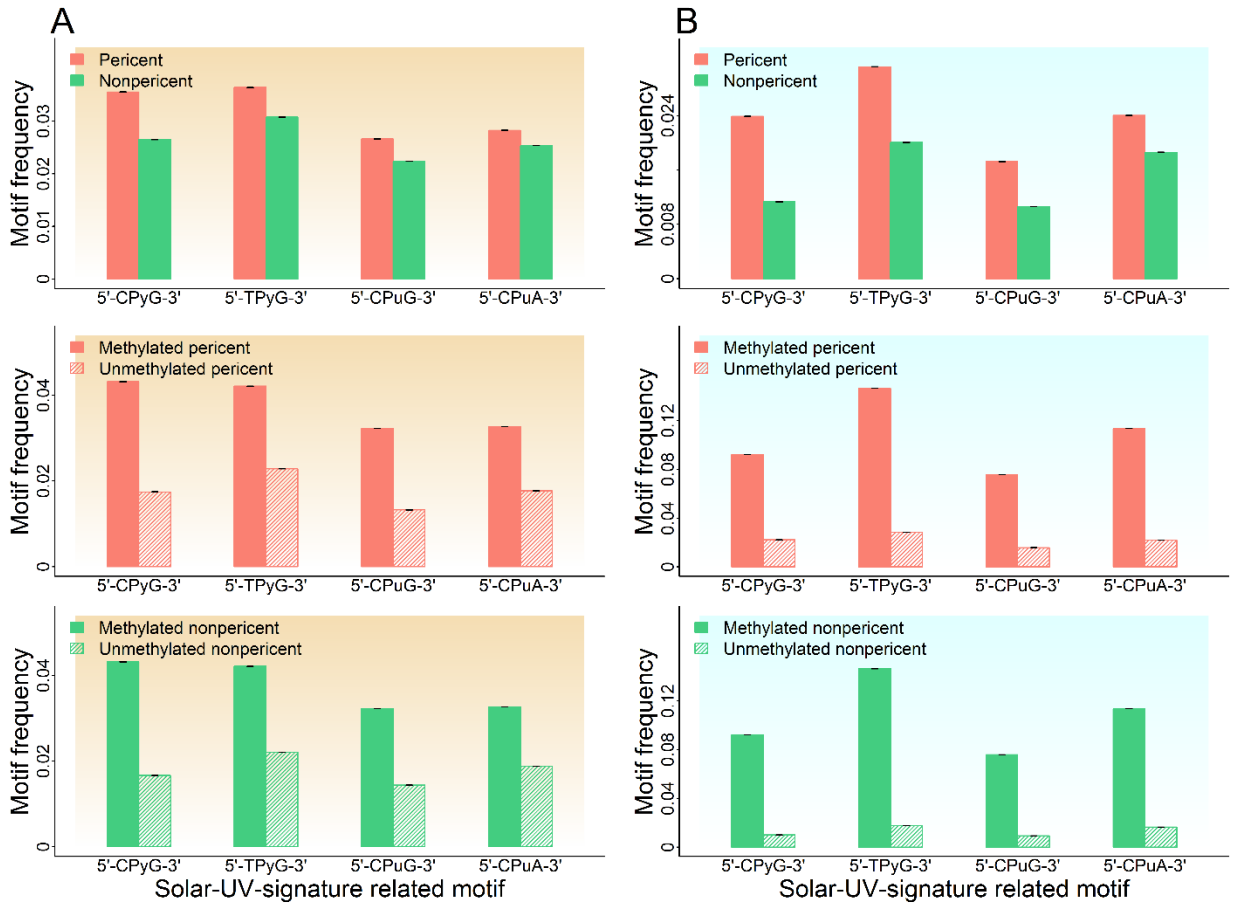


Figure S18. Frequencies of motifs related to solar-UV signature among SNPs from pericentromeric and non-pericentromeric regions under methylated and unmethylated conditions in (**A**) maize and (**B**) soybean. The top panel shows the frequencies of motifs with SNPs from pericentromeric regions and SNPs from non-pericentromeric regions. The middle panel shows the frequencies of motifs with SNPs from pericentromeric regions conditional on methylated and unmethylated regions. The bottom panel shows the frequencies of motifs with SNPs from non-pericentromeric regions conditonal on methylated and unmethylated regions. Each bar represents the average frequency of a specific motif over 100 maize accessions in (**A**) and 302 soybean accessions in (**B**). The black bar on the top illustrates the standardard error.

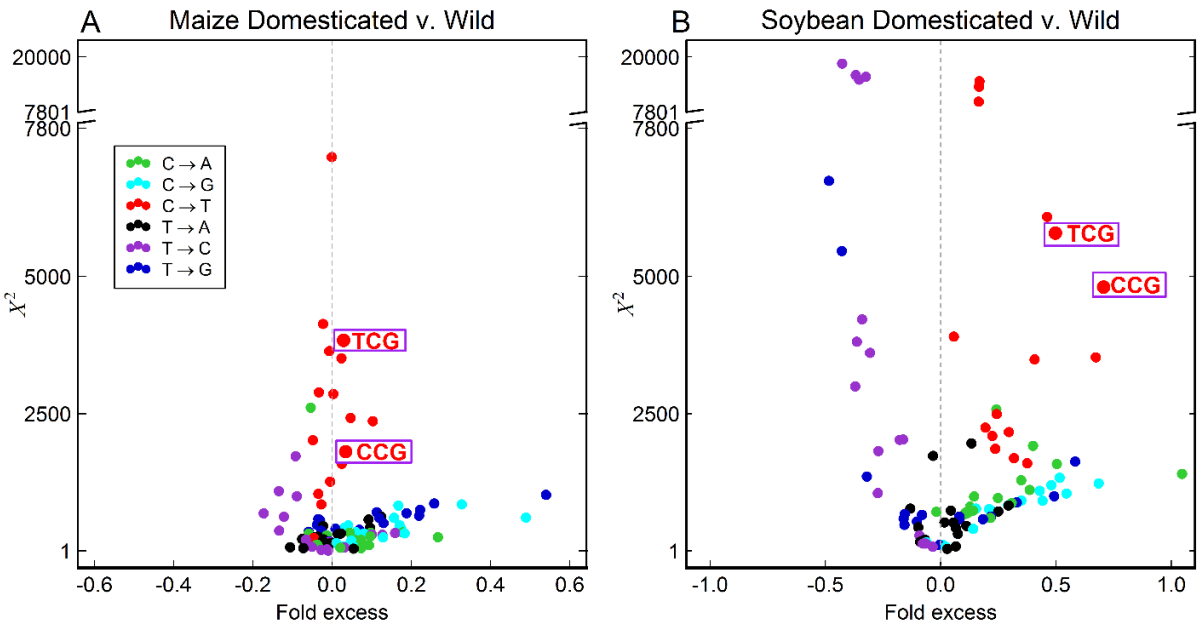


Figure S19. Enrichment test of mutations related to solar-UV signature with population-private SNPs. Compare the mutation frequency between domesticated accessions and wild accessions in (**A**) maize and (**B**) soybean. The x coordinate of each point indicates the fold frequency difference $(f_{PD}(m) - f_{PW}(m))/f_{PW}(m)$. The y coordinate indicates the Pearson's χ^2 value that measures the significance of the difference between $f_m(P_1)$ and $f_m(P_2)$. Outlier points are labeled with the ancestral state of the mutant nucleotide flanked by two neighboring bases, and the color of the points indicate the ancestral and derived alleles of the mutant site. The purple rectangle highlights the mutations related to solar-UV signature. Here TCG on the plot represents mutation 5′-TCG-3′→5′-TTG-3′ and its reverse complement 5′-CGA-3′→5′-CAA-3′, CCG represents mutation 5′-CCG-3′→5′-CTG-3′ and its reverse complement 5′-CGG-3′→5′-CAG-3′, and similarly for all the other dots on the plot.

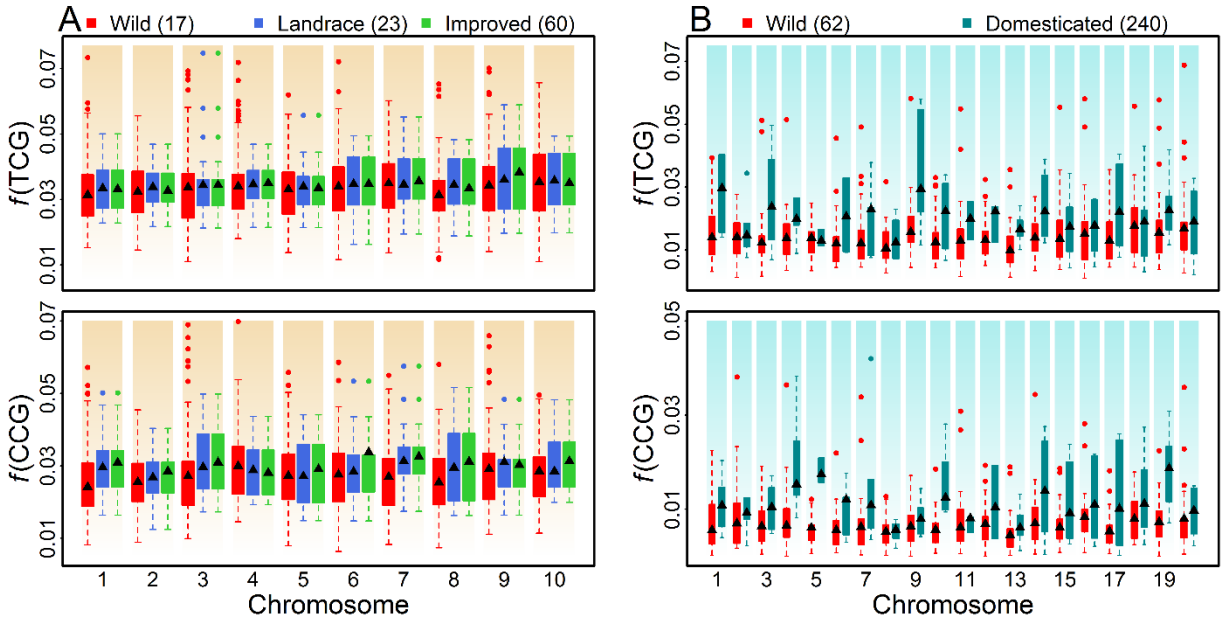


Figure S20. The distribution of $f(\mathrm{TCG})$ and $f(\mathrm{CCG})$ across bins of 1,000 consecutive population-private SNPs. (**A**) In maize, private SNP sets PW, PL and PI were analyzed. (**B**) In soybean, because of the small number of private SNPs in PL and PI, private SNP sets PD and PW were analyzed. Each private SNP set was partitioned into 1,000 consecutive SNP bins on each chromosomes that are not overlapped with each other. The frequency $f(\mathrm{TCG})$ for TCG→T mutation (5′-TCG-3′→5′-TTG-3′ and its reverse complement 5′-CGA-3′→5′-CAA-3′), and the frequency $f(\mathrm{CCG})$ for CCG→T mutation (5′-CCG-3′→5′-CTG-3′ and its reverse complement 5′-CGG-3′→5′-CAG-3′) of each bin were calculated and plotted. The black triangle within each box plot indicates the chromosome-wide frequency.

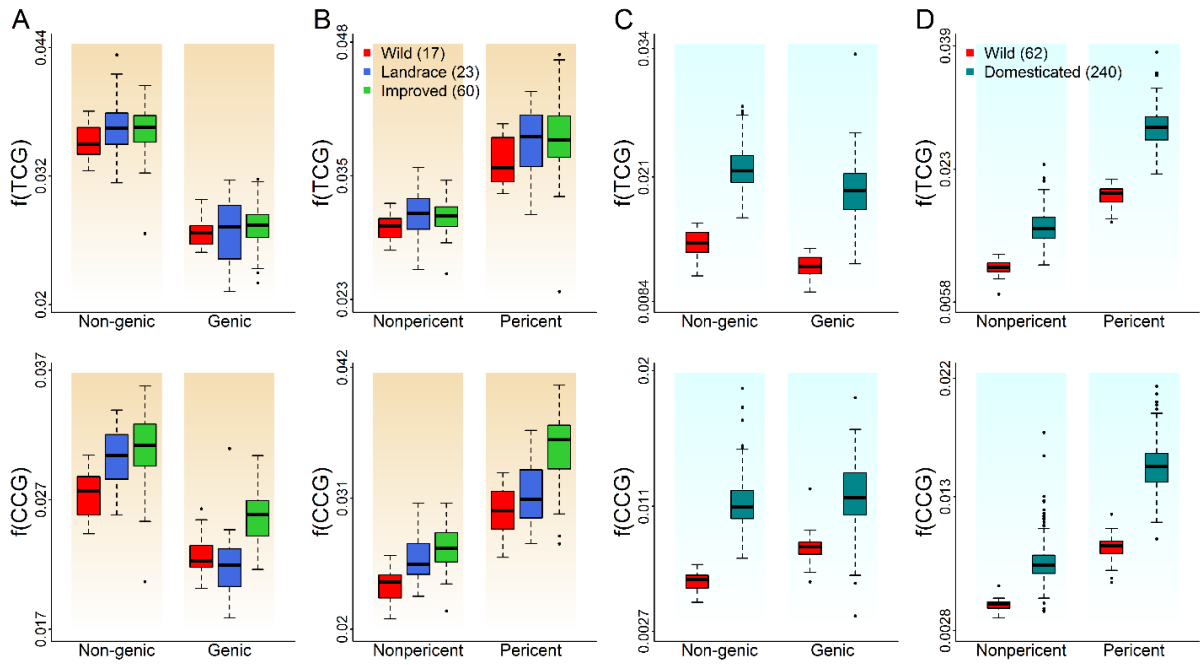


Figure S21. The distribution of $f(\mathrm{TCG})$ and $f(\mathrm{CCG})$ at different genomic regions. (**A**) In maize, compare $f(\mathrm{TCG})$ and $f(\mathrm{CCG})$ calculated with non-genic-private SNPs to those calculated with genic-private SNPs (red for wild accession, blue for landraces, and green for improved cultivars). (**B**) In maize, compare $f(\mathrm{TCG})$ and $f(\mathrm{CCG})$ calculated from pericentromeric-private SNPs to those calculated from non-pericentromeric-private SNPs. (**C**) In soybean, compare $f(\mathrm{TCG})$ and $f(\mathrm{CCG})$ calculated with non-genic- private SNPs to those calculated with genic-private SNPs (red for wild accession, turquoise for domesticated accession). (**D**) In soybean, compare $f(\mathrm{TCG})$ and $f(\mathrm{CCG})$ calculated from pericentromeric-private SNPs to those calculated from non-pericentromeric-private SNPs. $f(\mathrm{TCG})$ is the frequency of TCG→T mutation (5′-TCG-3′→5′-TTG-3′ and its reverse complement 5′-CGA-3′→5′-CAA-3′), and $f(\mathrm{CCG})$ is the frequency of CCG→T mutation (5′-CCG-3′→5′-CTG-3′ and its reverse complement 5′-CGG-3′→5′-CAG-3′).

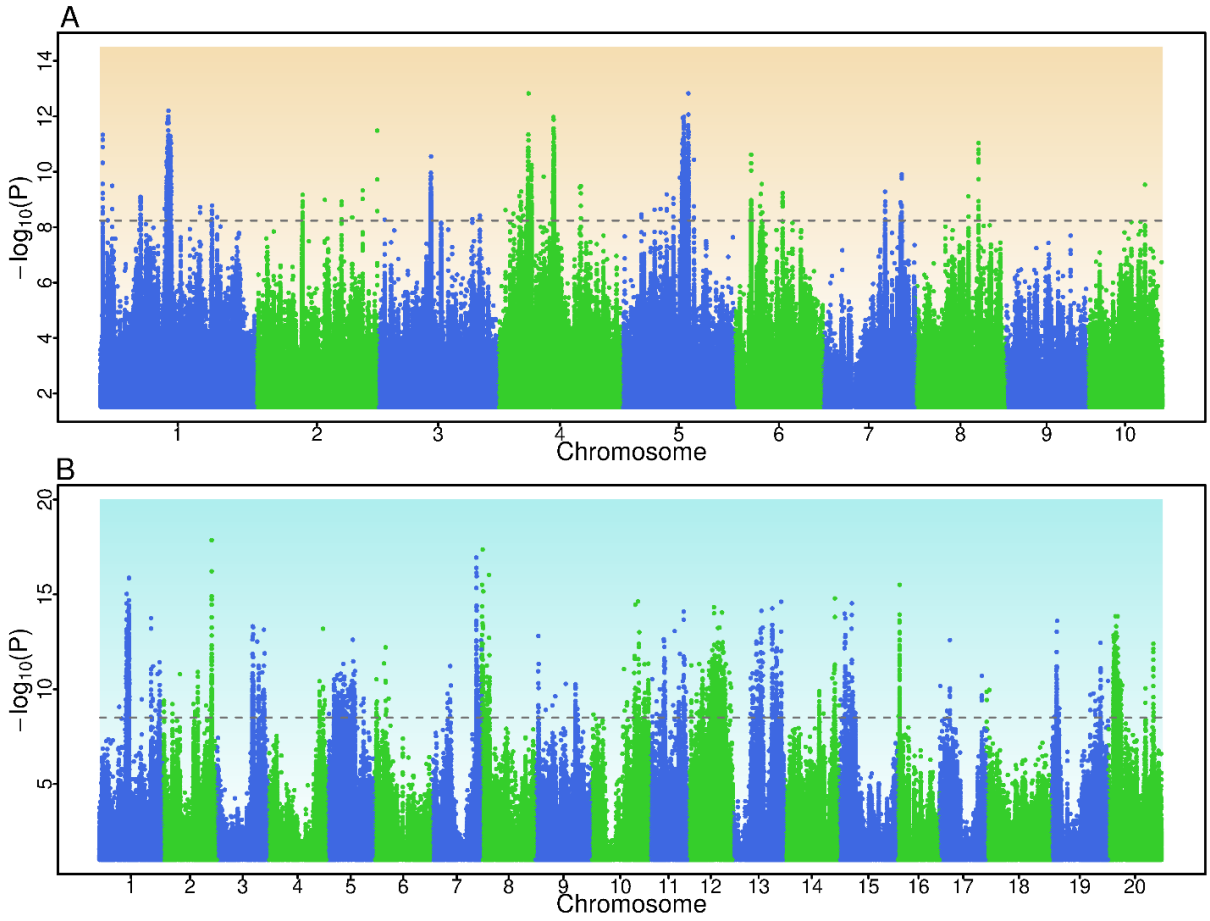


Figure S22. Genome-wide scan with a mixed model to identify genomic regions underlying base composition variation in (**A**) maize and (**B**) soybean. Manhattan plot shows the association signals detected by the mixed model between the genome-wide [AT] values across polymorphic sites in (**A**) 100 maize accessions and (**B**) 302 soybean accessions.

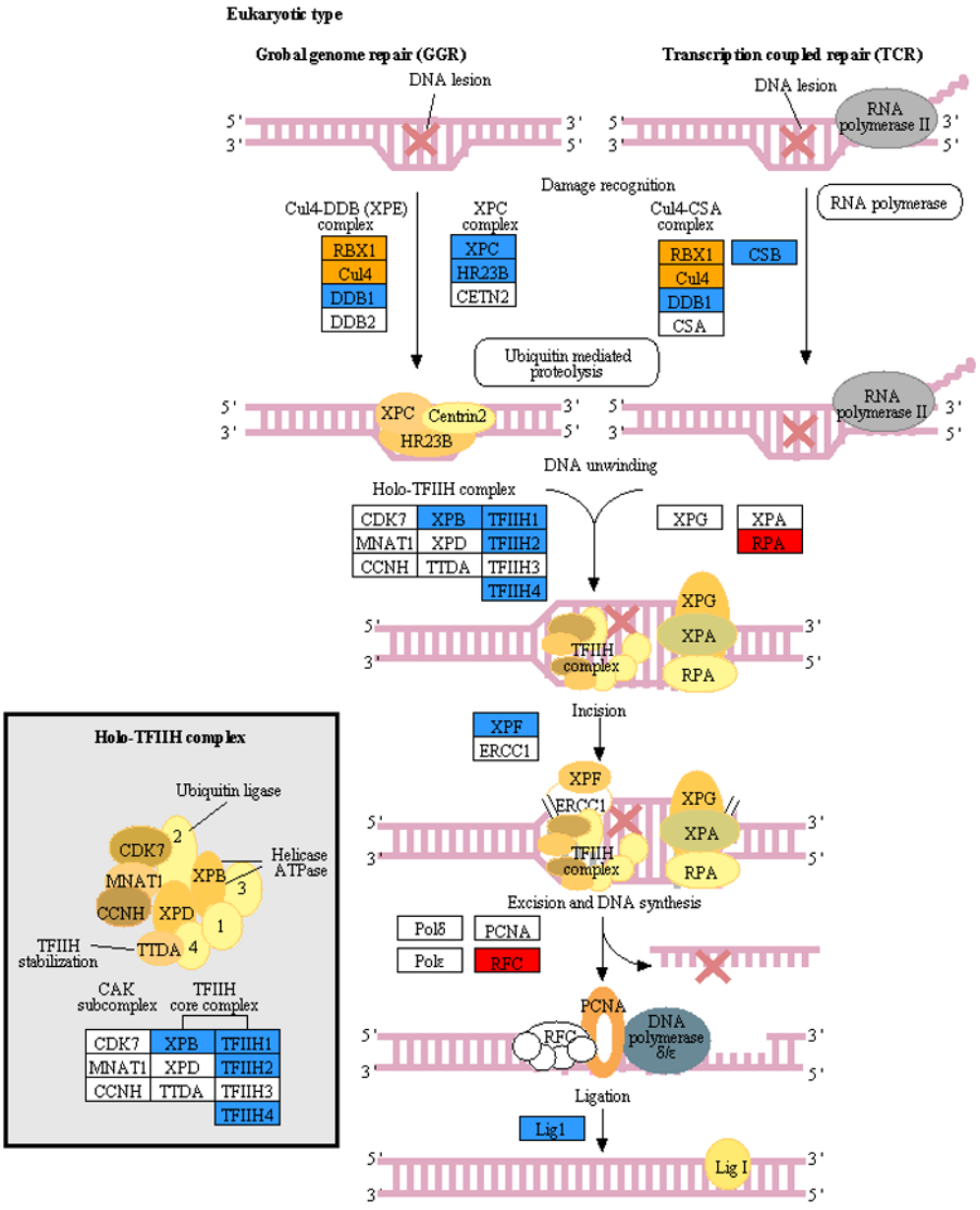


Figure S23. GWAS tagged genes in NER pathway. The pathway was obtained from KEGG (Nucleotide excision repair, ath:03420). Genes with orange box and blue box are located within 500kb from significantly associated SNPs in maize and soybean, respectively. And genes with red box are detected in both crops.

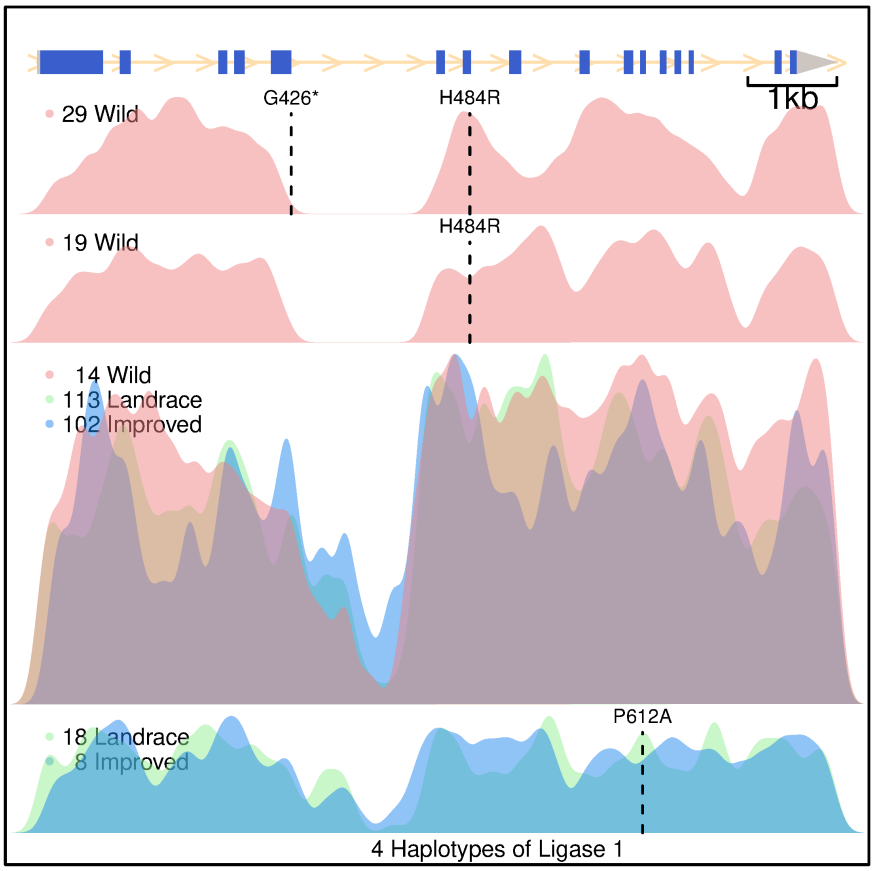


Figure S24. DNA short reads alignment reveals the structural variation in soybean *DNA ligase1* (*Lig1*). The coverage of mapped short reads was plotted to shown the indel in the 5th intron. Haplotypes of *Lig1* formed by 2 nonsynonymous SNPs and a 1.8kb indel among 302 accessions.

Table S1. UV-related genes are enriched near the associated loci in maize

Distance from signals (Mb)	Frequency of detected all encoded genes	Frequency of detected UV-related gene	*P*-value
0.50	1.8%	4.2%	0.002*
1.00	2.8%	5.4%	0.004*
1.50	4.0%	7.8%	0.001*
2.00	5.4%	9.6%	0.001*
2.50	6.7%	10.8%	0.002*

Table S2. UV-related genes located within 500kb from the associated SNPs in maize

Genes	Chr	Start	End	*Arabidopsis* orthologue	Alias	Function
Zm00001d030376	1	127,344,674	127,353,195	AT4G19130	*RPA1*	Replication factor-A protein 1-related
Zm00001d030381	1	127,553,498	127,556,739	AT1G12370	*UVR2*	Photolyase 1
Zm00001d005361	2	171,059,897	171,064,674	AT3G53570	*CLK2B*	CDC2-related kinase subfamily
Zm00001d007897	2	241,927,996	241,932,729	AT2G02760	*ATUBC2*	UBC2 ubiquiting-conjugating enzyme
Zm00001d042988	3	185,995,285	185,995,920	AT5G01310	*APTX*	Adenylylsulfate sulfohydrolase activity, involved in base exicision repair
Zm00001d049471	4	31,448,746	31,454,292	AT5G54260	*MRE11A*	DNA repair and meiotic recombination protein
Zm00001d049811	4	45,778,051	45,785,382	AT2G06510	*RPA1*	Encodes a homolog of Replication protein A
Zm00001d050085	4	64,882,472	64,888,748	AT5G46210	*CUL4*	Ubiqutin protein ligase activity, involved in DNA repair
Zm00001d050642	4	109,006,158	109,009,903	AT5G22750	*RAD5*	DNA/RNA helicase protein involves in DNA repair
Zm00001d051565	4	162,887,403	162,902,341	AT5G27740	*RFC3*	DNA repair, DNA-dependent DNA replication
Zm00001d051588	4	163,585,208	163,588,502	AT5G20570	*RBX1*	Subunit of Cul2-RING ubiquitin ligase complex, involved in nucleotide exicision repair
Zm00001d014813	5	64,068,899	64,100,088	AT5G40820	*ATR*	Encodes a Arabidopsis ortholog of the ATR protein kinase
Zm00001d015871	5	127,459,293	127,465,455	AT3G48750	*CDK2*	A-type cyclin-dependent kinase, involved in DNA repair
Zm00001d021607	7	157,927,203	157,934,056	AT1G05120	*RAD16*	ATP-binding protein, required for nucleotide exicision repair

Table S3. UV-related genes are enriched near the associated loci in soybean.

Distance from signals (Mb)	Frequency of detected all encoded genes	Frequency of detected UV-related gene	*P*-value
0.20	8.0%	11.2%	0.221
0.50	13.8%	20.6%	0.041*
1.00	22.1%	30.8%	0.029*
1.50	29.0%	43.0%	0.001*
2.00	33.6%	50.5%	0.001*

Table S4. UV-related genes located within 500kb from the associated SNPs in soybean

Genes	Chr	Start	End	*Arabidopsis* orthologue	Alias	Function
Glyma.01g081500	1	23,600,207	23,609,665	AT3G02540	*RAD23C*	Rad23 UV excision repair protein family
Glyma.01g204500	1	53,723,913	53,729,465	AT5G41150	*UVH1*	Restriction endonuclease, type II-like superfamily protein
Glyma.01g212300	1	54,381,204	54,383,602	AT1G12370	*UVR2*	Photolyase 1
Glyma.02g182000	2	31,149,538	31,150,558	AT5G61000	*ATPRA70D*	Replication factor-A protein 1-related
Glyma.03g196800	3	40,638,356	40,640,900	AT4G18590	*AtRPA14B*	ssDNA binding protein
Glyma.04g215000	4	48,672,908	48,674,463	AT1G77470	*RFC5*	DNA-dependent ATPase required for DNA replication and repair
Glyma.04g223500	4	49,411,690	49,415,117	AT2G38560	*TFIIS*	Transcript elongation factor IIS
Glyma.08g016400	8	1,297,174	1,305,157	AT5G41370	*XPB1*	Subunit of TFIIH. 3'->5' helicase
Glyma.08g088900	8	6,714,191	6,721,326	AT5G44740	*POLH*	Y-family DNA polymerase H
Glyma.10g156300	10	39,053,531	39,057,116	AT3G05210	*UVR7*	Nucleotide repair protein
Glyma.10g193000	10	42,527,991	42,540,681	AT1G55750	*AtTFB1-1*	Core TFIIH subunits
Glyma.11g038500	11	2,750,496	2,758,704	AT5G41150	*UVH1*	Restriction endonuclease, type II-like superfamily protein
Glyma.11g193100	11	26,629,471	26,638,425	AT1G08130	*LIG1*	DNA ligase 1
Glyma.11g211200	11	30,385,621	30,386,438	AT4G17020	*AtTFB2*	Core TFIIH subunits
Glyma.12g096100	12	8,101,497	8,111,369	AT5G22010	*RFC1*	Replication factor C1
Glyma.12g106200	12	9,701,540	9,702,031	AT4G21100	*DDB1B*	Damaged DNA binding protein 1B
Glyma.13g096800	13	21,172,838	21,179,791	AT3G02920	*RPA32B*	Replication protein A, subunit RPA32
Glyma.13g245700	13	35,464,732	35,471,257	AT5G28740		TPR-like superfamily protein
Glyma.15g055100	15	4,326,815	4,329,488	AT3G50360	*CEN2*	Centrin2
Glyma.15g068100	15	5,200,760	5,206,464	AT5G28740		TPR-like superfamily protein
Glyma.19g044000	19	6,474,307	6,487,275	AT1G16710	*HAC12*	Histone acetyltransferase of the CBP family 12
Glyma.19g129000	19	38,832,877	38,838,251	AT1G05055	*GTF2H2*	General transcription factor II H2

Table S5. Summary of 100 maize accessions.

Accession	Category	Species	Class
TIL01	Parviglumis	*Z. mays* ssp. *parviglumis*	TIL
TIL03	Parviglumis	*Z. mays* ssp. *parviglumis*	TIL
TIL04 (TIP-454)	Parviglumis	*Z. mays* ssp. *parviglumis*	TIL
TIL05	Parviglumis	*Z. mays* ssp. *parviglumis*	TIL
TIL06 (TIP-260)	Parviglumis	*Z. mays* ssp. *parviglumis*	TIL
TIL06 (TIP-496)	Parviglumis	*Z. mays* ssp. *parviglumis*	TIL
TIL07	Parviglumis	*Z. mays* ssp. *parviglumis*	TIL
TIL08	Mexicana	*Z. mays* ssp. m*exicana*	TIL
TIL09	Parviglumis	*Z. mays* ssp. *parviglumis*	TIL
TIL10	Parviglumis	*Z. mays* ssp. *parviglumis*	TIL
TIL11	Parviglumis	*Z. mays* ssp. *parviglumis*	TIL
TIL12	Parviglumis	*Z. mays* ssp. *parviglumis*	TIL
TIL14	Parviglumis	*Z. mays* ssp. *parviglumis*	TIL
TIL15	Parviglumis	*Z. mays* ssp. *parviglumis*	TIL
TIL16	Parviglumis	*Z. mays* ssp. *parviglumis*	TIL
TIL17	Parviglumis	*Z. mays* ssp. *parviglumis*	TIL
TIL25	Mexicana	*Z. mays* ssp. m*exicana*	TIL
MR01	Landrace	*Z. mays* ssp. *mays*	LRI
MR02	Landrace	*Z. mays* ssp. *mays*	LRI
MR03	Landrace	*Z. mays* ssp. *mays*	LRI
MR05	Landrace	*Z. mays* ssp. *mays*	LRI
MR06	Landrace	*Z. mays* ssp. *mays*	LRI
MR07	Landrace	*Z. mays* ssp. *mays*	LRI
MR08	Landrace	*Z. mays* ssp. *mays*	LRI
MR09	Landrace	*Z. mays* ssp. *mays*	LRI
MR10	Landrace	*Z. mays* ssp. *mays*	LRI
MR11	Landrace	*Z. mays* ssp. *mays*	LRI
MR12	Landrace	*Z. mays* ssp. *mays*	LRI
MR13	Landrace	*Z. mays* ssp. *mays*	LRI
MR14	Landrace	*Z. mays* ssp. *mays*	LRI
MR17	Landrace	*Z. mays* ssp. *mays*	LRI
MR18	Landrace	*Z. mays* ssp. *mays*	LRI
MR19	Landrace	*Z. mays* ssp. *mays*	LRI
MR20	Landrace	*Z. mays* ssp. *mays*	LRI
MR21	Landrace	*Z. mays* ssp. *mays*	LRI
MR22	Landrace	*Z. mays* ssp. *mays*	LRI
MR23	Landrace	*Z. mays* ssp. *mays*	LRI
MR24	Landrace	*Z. mays* ssp. *mays*	LRI
MR25	Landrace	*Z. mays* ssp. *mays*	LRI
MR26	Landrace	*Z. mays* ssp. *mays*	LRI
B73	Improved	*Z. mays* ssp. *mays*	SS
B97	Improved	*Z. mays* ssp. *mays*	NSS
CAU178	Improved	*Z. mays* ssp. *mays*	CAU
CAU478	Improved	*Z. mays* ssp. *mays*	CAU
CAU5003	Improved	*Z. mays* ssp. *mays*	CAU
CAUCHANG72	Improved	*Z. mays* ssp. *mays*	CAU
CAUMO17	Improved	*Z. mays* ssp. *mays*	NSS
CAUZHENG58	Improved	*Z. mays* ssp. *mays*	CAU
CML103	Improved	*Z. mays* ssp. *mays*	TS
CML133	Improved	*Z. mays* ssp. *mays*	CML
CML192	Improved	*Z. mays* ssp. *mays*	CML

Table S5 Continued

Accession	Category	Species	Class
CML202	Improved	*Z. mays* ssp. *mays*	CML
CML206	Improved	*Z. mays* ssp. *mays*	CML
CML228	Improved	*Z. mays* ssp. *mays*	TS
CML247	Improved	*Z. mays* ssp. *mays*	TS
CML277	Improved	*Z. mays* ssp. *mays*	TS
CML312SR	Improved	*Z. mays* ssp. *mays*	CML
CML322	Improved	*Z. mays* ssp. *mays*	TS
CML330	Improved	*Z. mays* ssp. *mays*	CML
CML333	Improved	*Z. mays* ssp. *mays*	TS
CML341	Improved	*Z. mays* ssp. *mays*	TS
CML411	Improved	*Z. mays* ssp. *mays*	CML
CML418	Improved	*Z. mays* ssp. *mays*	CML
CML479	Improved	*Z. mays* ssp. *mays*	CML
CML504	Improved	*Z. mays* ssp. *mays*	CML
CML505	Improved	*Z. mays* ssp. *mays*	CML
CML511	Improved	*Z. mays* ssp. *mays*	CML
CML52	Improved	*Z. mays* ssp. *mays*	TS
CML69	Improved	*Z. mays* ssp. *mays*	TS
CML84	Improved	*Z. mays* ssp. *mays*	CML
CML85	Improved	*Z. mays* ssp. *mays*	CML
CML96	Improved	*Z. mays* ssp. *mays*	CML
CML99	Improved	*Z. mays* ssp. *mays*	CML
H16	Improved	*Z. mays* ssp. *mays*	NotAssigned
HP301	Improved	*Z. mays* ssp. *mays*	POPCORN
IL14H	Improved	*Z. mays* ssp. *mays*	SWEET
KI11	Improved	*Z. mays* ssp. *mays*	TS
KI3	Improved	*Z. mays* ssp. *mays*	TS
KY21	Improved	*Z. mays* ssp. *mays*	NSS
M162W	Improved	*Z. mays* ssp. *mays*	NSS
M37W	Improved	*Z. mays* ssp. *mays*	MIXED
MO17	Improved	*Z. mays* ssp. *mays*	NSS
MO18W	Improved	*Z. mays* ssp. *mays*	MIXED
MS71	Improved	*Z. mays* ssp. *mays*	NSS
NC350	Improved	*Z. mays* ssp. *mays*	TS
NC358	Improved	*Z. mays* ssp. *mays*	TS
OH43	Improved	*Z. mays* ssp. *mays*	NSS
OH7B	Improved	*Z. mays* ssp. *mays*	NSS
P1	Improved	*Z. mays* ssp. *mays*	NotAssigned
P39	Improved	*Z. mays* ssp. *mays*	SWEET
TX303	Improved	*Z. mays* ssp. *mays*	MIXED
TZI8	Improved	*Z. mays* ssp. *mays*	TS
VL0512447	Improved	*Z. mays* ssp. *mays*	Chinese Tropical
VL05128	Improved	*Z. mays* ssp. *mays*	Chinese Tropical
VL054178	Improved	*Z. mays* ssp. *mays*	Chinese Tropical
VL05610	Improved	*Z. mays* ssp. *mays*	Chinese Tropical
VL056883	Improved	*Z. mays* ssp. *mays*	Chinese Tropical
VL062784	Improved	*Z. mays* ssp. *mays*	Chinese Tropical
W22	Improved	*Z. mays* ssp. *mays*	NSS
W64A	Improved	*Z. mays* ssp. *mays*	NSS

Table S6. Summary of 302 soybean accessions.

Accession	Category	Species	PI CGN# & Name
IGDB-001	G. soja	*G. soja* Siebold & Zucc.	ZJ-ZY020
IGDB-002	G. soja	*G. soja* Siebold & Zucc.	ZJ-YJ086
IGDB-003	G. soja	*G. soja* Siebold & Zucc.	ZJ-Y314
IGDB-004	G. soja	*G. soja* Siebold & Zucc.	ZJ-Y217
IGDB-005	G. soja	*G. soja* Siebold & Zucc.	ZJ-Y200
IGDB-006	G. soja	*G. soja* Siebold & Zucc.	ZJ-Y191
IGDB-007	G. soja	*G. soja* Siebold & Zucc.	ZJ-Y188
IGDB-008	G. soja	*G. soja* Siebold & Zucc.	ZJ-Y108
IGDB-009	G. soja	*G. soja* Siebold & Zucc.	ZJ-YJ038
IGDB-010	G. soja	*G. soja* Siebold & Zucc.	ZJ-Y282
IGDB-011	G. soja	*G. soja* Siebold & Zucc.	ZJ-Y2300-1
IGDB-012	G. soja	*G. soja* Siebold & Zucc.	ZJ-Y155
IGDB-013	G. soja	*G. soja* Siebold & Zucc.	PI 597461C
IGDB-014	G. soja	*G. soja* Siebold & Zucc.	PI 597461A
IGDB-015	G. soja	*G. soja* Siebold & Zucc.	PI 597459D
IGDB-016	G. soja	*G. soja* Siebold & Zucc.	PI 597459C
IGDB-017	G. soja	*G. soja* Siebold & Zucc.	PI 593983
IGDB-018	G. soja	*G. soja* Siebold & Zucc.	PI 578357
IGDB-019	G. soja	*G. soja* Siebold & Zucc.	PI 578341
IGDB-020	G. soja	*G. soja* Siebold & Zucc.	PI 562565
IGDB-021	G. soja	*G. soja* Siebold & Zucc.	PI 562559
IGDB-022	G. soja	*G. soja* Siebold & Zucc.	PI 549046
IGDB-023	G. soja	*G. soja* Siebold & Zucc.	PI 547831
IGDB-024	G. soja	*G. soja* Siebold & Zucc.	PI 522228
IGDB-025	G. soja	*G. soja* Siebold & Zucc.	PI 522226
IGDB-026	G. soja	*G. soja* Siebold & Zucc.	PI 522216
IGDB-027	G. soja	*G. soja* Siebold & Zucc.	PI 522182B
IGDB-028	G. soja	*G. soja* Siebold & Zucc.	PI 507662
IGDB-029	G. soja	*G. soja* Siebold & Zucc.	PI 504286
IGDB-030	G. soja	*G. soja* Siebold & Zucc.	PI 483465
IGDB-031	G. soja	*G. soja* Siebold & Zucc.	PI 483464A
IGDB-032	G. soja	*G. soja* Siebold & Zucc.	PI 483460B
IGDB-033	G. soja	*G. soja* Siebold & Zucc.	PI 479769
IGDB-034	G. soja	*G. soja* Siebold & Zucc.	PI 479752
IGDB-035	G. soja	*G. soja* Siebold & Zucc.	PI 468916
IGDB-036	G. soja	*G. soja* Siebold & Zucc.	PI 468400A
IGDB-037	G. soja	*G. soja* Siebold & Zucc.	PI 464935
IGDB-038	G. soja	*G. soja* Siebold & Zucc.	PI 464929B
IGDB-039	G. soja	*G. soja* Siebold & Zucc.	PI 464929A
IGDB-040	G. soja	*G. soja* Siebold & Zucc.	PI 464927A
IGDB-041	G. soja	*G. soja* Siebold & Zucc.	PI 458538
IGDB-042	G. soja	*G. soja* Siebold & Zucc.	PI 458536
IGDB-043	G. soja	*G. soja* Siebold & Zucc.	PI 458535
IGDB-044	G. soja	*G. soja* Siebold & Zucc.	PI 447004
IGDB-045	G. soja	*G. soja* Siebold & Zucc.	PI 424096
IGDB-046	G. soja	*G. soja* Siebold & Zucc.	PI 423991
IGDB-047	G. soja	*G. soja* Siebold & Zucc.	PI 407301
IGDB-048	G. soja	*G. soja* Siebold & Zucc.	PI 407288
IGDB-049	G. soja	*G. soja* Siebold & Zucc.	PI 407285
IGDB-050	G. soja	*G. soja* Siebold & Zucc.	PI 407275
IGDB-051	G. soja	*G. soja* Siebold & Zucc.	PI 407246

Table S6 Continued

Accession	Category	Species	PI CGN# & Name
IGDB-052	G. soja	*G. soja* Siebold & Zucc.	PI 407197
IGDB-053	G. soja	*G. soja* Siebold & Zucc.	PI 407170
IGDB-054	G. soja	*G. soja* Siebold & Zucc.	PI 407131
IGDB-055	G. soja	*G. soja* Siebold & Zucc.	PI 407027
IGDB-056	G. soja	*G. soja* Siebold & Zucc.	PI 393551
IGDB-057	G. soja	*G. soja* Siebold & Zucc.	PI 378692
IGDB-058	G. soja	*G. soja* Siebold & Zucc.	PI 366123
IGDB-059	G. soja	*G. soja* Siebold & Zucc.	PI 366121
IGDB-060	G. soja	*G. soja* Siebold & Zucc.	PI 366120
IGDB-061	G. soja	*G. soja* Siebold & Zucc.	PI 339871A
IGDB-062	G. soja	*G. soja* Siebold & Zucc.	PI 326582A
IGDB-063	Landrace	*G. max* (L.) Merr.	Yu Shi Dou
IGDB-064	Landrace	*G. max* (L.) Merr.	Yu Jiang Wu Yue Niu Mao Huang
IGDB-065	Landrace	*G. max* (L.) Merr.	You Pi Zhi Hei Dou
IGDB-066	Landrace	*G. max* (L.) Merr.	Yi Zheng Da Li Huang Dou
IGDB-067	Landrace	*G. max* (L.) Merr.	Xin Xian Xiao Huang Dou
IGDB-068	Landrace	*G. max* (L.) Merr.	Xiao Mi Dou
IGDB-069	Landrace	*G. max* (L.) Merr.	Xiao Huang Dou
IGDB-070*	Landrace	*G. max* (L.) Merr.	Xiao Bai Qi
IGDB-071	Landrace	*G. max* (L.) Merr.	Xiao Bai Dou
IGDB-072	Landrace	*G. max* (L.) Merr.	Xiang Dou No.4
IGDB-073	Landrace	*G. max* (L.) Merr.	Xia Men Teng Zai Dou
IGDB-074	Landrace	*G. max* (L.) Merr.	Xia Hei Dou
IGDB-075	Landrace	*G. max* (L.) Merr.	Tong an Zi Hong Dou
IGDB-076	Landrace	*G. max* (L.) Merr.	Tian E Dan
IGDB-077	Landrace	*G. max* (L.) Merr.	Tai Xin Niu Mao Huang Yi
IGDB-078	Landrace	*G. max* (L.) Merr.	Sha Xin Dou
IGDB-079	Landrace	*G. max* (L.) Merr.	Sha Xian Wu Dou
IGDB-080	Landrace	*G. max* (L.) Merr.	Sha Xian Qin Dou
IGDB-081	Landrace	*G. max* (L.) Merr.	Qing Dou
IGDB-082	Landrace	*G. max* (L.) Merr.	PI Xian Nian Zhuang Liu Yue Xian
IGDB-083	Landrace	*G. max* (L.) Merr.	PI Xian Da Zi Huo Cao
IGDB-084	Landrace	*G. max* (L.) Merr.	PI 89138
IGDB-085	Landrace	*G. max* (L.) Merr.	PI 88479
IGDB-086	Landrace	*G. max* (L.) Merr.	PI 86024
IGDB-087	Landrace	*G. max* (L.) Merr.	PI 84987A
IGDB-088	Landrace	*G. max* (L.) Merr.	PI 84987
IGDB-089	Landrace	*G. max* (L.) Merr.	PI 84631
IGDB-090	Landrace	*G. max* (L.) Merr.	PI 83945-3
IGDB-091	Landrace	*G. max* (L.) Merr.	PI 80837
IGDB-092	Landrace	*G. max* (L.) Merr.	PI 80822
IGDB-093	Improved	*G. max* (L.) Merr.	PI 634883
IGDB-094	Landrace	*G. max* (L.) Merr.	PI 603756
IGDB-095	Landrace	*G. max* (L.) Merr.	PI 603675
IGDB-096	Landrace	*G. max* (L.) Merr.	PI 603596
IGDB-097	Landrace	*G. max* (L.) Merr.	PI 603516
IGDB-098	Landrace	*G. max* (L.) Merr.	PI 603424A
IGDB-099	Landrace	*G. max* (L.) Merr.	PI 603420
IGDB-100*	Landrace	*G. max* (L.) Merr.	PI 603384
IGDB-101	Landrace	*G. max* (L.) Merr.	PI 603357
IGDB-102	Landrace	*G. max* (L.) Merr.	PI 603336

Table S6 Continued

Accession	Category	Species	PI CGN# & Name
IGDB-103	Landrace	*G. max* (L.) Merr.	PI 603318
IGDB-104	Landrace	*G. max* (L.) Merr.	PI 602991
IGDB-105	Landrace	*G. max* (L.) Merr.	PI 594788
IGDB-106	Landrace	*G. max* (L.) Merr.	PI 594777
IGDB-107	Landrace	*G. max* (L.) Merr.	PI 594629
IGDB-108	Landrace	*G. max* (L.) Merr.	PI 594615
IGDB-109	Landrace	*G. max* (L.) Merr.	PI 594579
IGDB-110	Landrace	*G. max* (L.) Merr.	PI 594451
IGDB-111	Landrace	*G. max* (L.) Merr.	PI 594301
IGDB-112	Improved	*G. max* (L.) Merr.	PI 591511
IGDB-113	Improved	*G. max* (L.) Merr.	PI 591495
IGDB-114	Landrace	*G. max* (L.) Merr.	PI 588053A
IGDB-115	Landrace	*G. max* (L.) Merr.	PI 587848
IGDB-116	Landrace	*G. max* (L.) Merr.	PI 587752
IGDB-117	Landrace	*G. max* (L.) Merr.	PI 587666
IGDB-118	Landrace	*G. max* (L.) Merr.	PI 587552
IGDB-119	Landrace	*G. max* (L.) Merr.	PI 578457A
IGDB-120	Landrace	*G. max* (L.) Merr.	PI 567525
IGDB-121	Landrace	*G. max* (L.) Merr.	PI 567503
IGDB-122	Landrace	*G. max* (L.) Merr.	PI 567395
IGDB-123	Landrace	*G. max* (L.) Merr.	PI 567364
IGDB-124	Landrace	*G. max* (L.) Merr.	PI 567298
IGDB-125	Landrace	*G. max* (L.) Merr.	PI 567293
IGDB-126	Landrace	*G. max* (L.) Merr.	PI 567258
IGDB-127	Landrace	*G. max* (L.) Merr.	PI 567189A
IGDB-128	Landrace	*G. max* (L.) Merr.	PI 567071A
IGDB-129	Landrace	*G. max* (L.) Merr.	PI 548488
IGDB-130	Landrace	*G. max* (L.) Merr.	PI 548485
IGDB-131	Improved	*G. max* (L.) Merr.	PI 548477
IGDB-132	Landrace	*G. max* (L.) Merr.	PI 548456
IGDB-133	Landrace	*G. max* (L.) Merr.	PI 548445
IGDB-134	Landrace	*G. max* (L.) Merr.	PI 548417
IGDB-135	Landrace	*G. max* (L.) Merr.	PI 548406
IGDB-136	Landrace	*G. max* (L.) Merr.	PI 548402
IGDB-137	Landrace	*G. max* (L.) Merr.	PI 548391
IGDB-138	Landrace	*G. max* (L.) Merr.	PI 548382
IGDB-139	Landrace	*G. max* (L.) Merr.	PI 548379
IGDB-140	Improved	*G. max* (L.) Merr.	PI 548362
IGDB-141	Landrace	*G. max* (L.) Merr.	PI 548348
IGDB-142	Landrace	*G. max* (L.) Merr.	PI 548342
IGDB-143	Improved	*G. max* (L.) Merr.	PI 548311
IGDB-144	Landrace	*G. max* (L.) Merr.	PI 548298
IGDB-145	Improved	*G. max* (L.) Merr.	PI 548190
IGDB-146	Improved	*G. max* (L.) Merr.	PI 548182
IGDB-147	Improved	*G. max* (L.) Merr.	PI 547562
IGDB-148	Landrace	*G. max* (L.) Merr.	PI 507355
IGDB-149	Landrace	*G. max* (L.) Merr.	PI 467343
IGDB-150	Landrace	*G. max* (L.) Merr.	PI 438498
IGDB-151	Landrace	*G. max* (L.) Merr.	PI 437944
IGDB-152	Landrace	*G. max* (L.) Merr.	PI 437679
IGDB-153	Landrace	*G. max* (L.) Merr.	PI 437654

Table S6 Continued

Accession	Category	Species	PI CGN# & Name
IGDB-154	Landrace	*G. max* (L.) Merr.	PI 437653
IGDB-155	Landrace	*G. max* (L.) Merr.	PI 437321
IGDB-156	Landrace	*G. max* (L.) Merr.	PI 424391
IGDB-157	Landrace	*G. max* (L.) Merr.	PI 423967
IGDB-158	Landrace	*G. max* (L.) Merr.	PI 423954
IGDB-159	Landrace	*G. max* (L.) Merr.	PI 417398
IGDB-160	Landrace	*G. max* (L.) Merr.	PI 416971
IGDB-161	Landrace	*G. max* (L.) Merr.	PI 416890
IGDB-162	Landrace	*G. max* (L.) Merr.	PI 407849
IGDB-163	Landrace	*G. max* (L.) Merr.	PI 407801
IGDB-164	Landrace	*G. max* (L.) Merr.	PI 407716
IGDB-165	Landrace	*G. max* (L.) Merr.	PI 404182
IGDB-166	Landrace	*G. max* (L.) Merr.	PI 399043
IGDB-167	Landrace	*G. max* (L.) Merr.	PI 398296
IGDB-168	Landrace	*G. max* (L.) Merr.	PI 339734
IGDB-169	Landrace	*G. max* (L.) Merr.	PI 323576
IGDB-170	Landrace	*G. max* (L.) Merr.	PI 317336
IGDB-171	Landrace	*G. max* (L.) Merr.	PI 317334A
IGDB-172	Landrace	*G. max* (L.) Merr.	PI 253658B
IGDB-173	Landrace	*G. max* (L.) Merr.	PI 243541
IGDB-174	Landrace	*G. max* (L.) Merr.	PI 196166
IGDB-175	Landrace	*G. max* (L.) Merr.	PI 157421
IGDB-176	Landrace	*G. max* (L.) Merr.	PI 153262
IGDB-177	Landrace	*G. max* (L.) Merr.	Pei Xian Xiao You Dou
IGDB-178	Landrace	*G. max* (L.) Merr.	Nian Shi Huang Dou
IGDB-179	Landrace	*G. max* (L.) Merr.	Ni Dou
IGDB-180	Landrace	*G. max* (L.) Merr.	Ni Ding Hua Mei Dou
IGDB-181*	Landrace	*G. max* (L.) Merr.	Nan Guan Xiao PI Qing
IGDB-182	Landrace	*G. max* (L.) Merr.	Long quan Da Dou
IGDB-183	Landrace	*G. max* (L.) Merr.	Jin Shan Cha Zhu Shi Dou
IGDB-184	Landrace	*G. max* (L.) Merr.	Jin Huang No.35
IGDB-185	Landrace	*G. max* (L.) Merr.	Ji Shan De Da Li Hei Dou
IGDB-186*	Landrace	*G. max* (L.) Merr.	Hu Pi Dou
IGDB-187	Landrace	*G. max* (L.) Merr.	Hong Zhu Dou
IGDB-188	Landrace	*G. max* (L.) Merr.	Hong Hu Liu Yue Bao
IGDB-189	Landrace	*G. max* (L.) Merr.	Hei Wa Shi Dou
IGDB-190	Landrace	*G. max* (L.) Merr.	Hei He Xiao Huang Dou
IGDB-191	Landrace	*G. max* (L.) Merr.	He Dou
IGDB-192	Landrace	*G. max* (L.) Merr.	Guang Rao Da Qing Dou
IGDB-193	Improved	*G. max* (L.) Merr.	FC 33243
IGDB-194	Landrace	*G. max* (L.) Merr.	Dong Shan Bai Ma Dou
IGDB-195	Landrace	*G. max* (L.) Merr.	Dai Mi Dou
IGDB-196	Landrace	*G. max* (L.) Merr.	Da Tun Xiao Hei Dou
IGDB-197	Landrace	*G. max* (L.) Merr.	Da Qing Ren
IGDB-198	Landrace	*G. max* (L.) Merr.	Da Li Huang
IGDB-199	Landrace	*G. max* (L.) Merr.	Cu Dou
IGDB-200	Landrace	*G. max* (L.) Merr.	Bin Hai Da huang Ke Zi Jia
IGDB-201	Improved	*G. max* (L.) Merr.	Beijing-IGDB-1
IGDB-202	Landrace	*G. max* (L.) Merr.	Bai Mao Dou
IGDB-203	Landrace	*G. max* (L.) Merr.	Bai Lu Dou
IGDB-204	Improved	*G. max* (L.) Merr.	Zhong Huang No.50

Table S6 Continued

Accession	Category	Species	PI CGN# & Name
IGDB-205	Improved	*G. max* (L.) Merr.	Zhong Huang No.40
IGDB-206	Improved	*G. max* (L.) Merr.	Zhong Huang No.38
IGDB-207	Improved	*G. max* (L.) Merr.	Zhong Huang No.35
IGDB-208	Improved	*G. max* (L.) Merr.	Zhong Huang No.31
IGDB-209	Improved	*G. max* (L.) Merr.	Zhong Huang No.14
IGDB-210	Improved	*G. max* (L.) Merr.	Zhong Huang No.13
IGDB-211	Improved	*G. max* (L.) Merr.	Xi Zang Da Dou No.20
IGDB-212	Improved	*G. max* (L.) Merr.	Wei No.6823
IGDB-213	Improved	*G. max* (L.) Merr.	Tie Feng No.22
IGDB-214	Improved	*G. max* (L.) Merr.	Tai Wan No.1
IGDB-215	Improved	*G. max* (L.) Merr.	Su Nong No.33
IGDB-216	Improved	*G. max* (L.) Merr.	Su Nong No.25
IGDB-217	Improved	*G. max* (L.) Merr.	Su Nong No.14
IGDB-218	Improved	*G. max* (L.) Merr.	Su Nong No.10
IGDB-219	Improved	*G. max* (L.) Merr.	Shu Xian No.205
IGDB-220	Improved	*G. max* (L.) Merr.	Sheng Dou No.9
IGDB-221	Improved	*G. max* (L.) Merr.	Shen Li No.3
IGDB-222	Improved	*G. max* (L.) Merr.	Harbin 91-6065
IGDB-223	Improved	*G. max* (L.) Merr.	PI 591541
IGDB-224	Improved	*G. max* (L.) Merr.	PI 591435
IGDB-225	Improved	*G. max* (L.) Merr.	PI 591433
IGDB-226	Improved	*G. max* (L.) Merr.	PI 591432
IGDB-227	Improved	*G. max* (L.) Merr.	PI 591431
IGDB-228	Improved	*G. max* (L.) Merr.	PI 553047
IGDB-229	Improved	*G. max* (L.) Merr.	PI 548985
IGDB-230	Improved	*G. max* (L.) Merr.	PI 548657
IGDB-231	Improved	*G. max* (L.) Merr.	PI 548644
IGDB-232	Improved	*G. max* (L.) Merr.	PI 548643
IGDB-233	Improved	*G. max* (L.) Merr.	PI 548638
IGDB-234	Improved	*G. max* (L.) Merr.	PI 548634
IGDB-235	Improved	*G. max* (L.) Merr.	PI 548631
IGDB-236	Improved	*G. max* (L.) Merr.	PI 548604
IGDB-237	Improved	*G. max* (L.) Merr.	PI 548603
IGDB-238	Improved	*G. max* (L.) Merr.	PI 548593
IGDB-239	Improved	*G. max* (L.) Merr.	PI 548573
IGDB-240	Improved	*G. max* (L.) Merr.	PI 548565
IGDB-241	Improved	*G. max* (L.) Merr.	PI 548540
IGDB-242	Improved	*G. max* (L.) Merr.	PI 548524
IGDB-243	Improved	*G. max* (L.) Merr.	PI 548520
IGDB-244	Improved	*G. max* (L.) Merr.	PI 548512
IGDB-245	Improved	*G. max* (L.) Merr.	PI 547779
IGDB-246	Improved	*G. max* (L.) Merr.	PI 547716
IGDB-247	Improved	*G. max* (L.) Merr.	PI 547690
IGDB-248	Improved	*G. max* (L.) Merr.	PI 547686
IGDB-249	Improved	*G. max* (L.) Merr.	PI 547680
IGDB-250	Improved	*G. max* (L.) Merr.	PI 547488
IGDB-251	Improved	*G. max* (L.) Merr.	PI 547460
IGDB-252	Improved	*G. max* (L.) Merr.	PI 547459
IGDB-253	Improved	*G. max* (L.) Merr.	PI 547409
IGDB-254	Improved	*G. max* (L.) Merr.	PI 546044
IGDB-255	Improved	*G. max* (L.) Merr.	PI 542403

Table S6 Continued

Accession	Category	Species	PI CGN# & Name
IGDB-256	Improved	*G. max* (L.) Merr.	PI 540552
IGDB-257	Improved	*G. max* (L.) Merr.	PI 536635
IGDB-258	Improved	*G. max* (L.) Merr.	PI 533655
IGDB-259	Improved	*G. max* (L.) Merr.	PI 533602
IGDB-260	Improved	*G. max* (L.) Merr.	PI 518750
IGDB-261	Improved	*G. max* (L.) Merr.	Jilin 21
IGDB-262	Improved	*G. max* (L.) Merr.	PI 518664
IGDB-263	Improved	*G. max* (L.) Merr.	PI 515961
IGDB-264	Improved	*G. max* (L.) Merr.	PI 513382
IGDB-265	Improved	*G. max* (L.) Merr.	PI 508266
IGDB-266	Improved	*G. max* (L.) Merr.	PI 508083
IGDB-267	Improved	*G. max* (L.) Merr.	Nan Nong Cai Dou No.1
IGDB-268	Improved	*G. max* (L.) Merr.	Lu Dou No.11
IGDB-269	Improved	*G. max* (L.) Merr.	Liao Dou No.3
IGDB-270	Improved	*G. max* (L.) Merr.	Liao Dou No.21
IGDB-271	Improved	*G. max* (L.) Merr.	Liao Dou No.17
IGDB-272	Improved	*G. max* (L.) Merr.	Liao Dou No.15
IGDB-273	Improved	*G. max* (L.) Merr.	Liao Dou No.11
IGDB-274	Improved	*G. max* (L.) Merr.	Jiu Nong No.30
IGDB-275	Improved	*G. max* (L.) Merr.	Jin Da No.75
IGDB-276	Improved	*G. max* (L.) Merr.	Jin Da No.73
IGDB-277	Improved	*G. max* (L.) Merr.	Jin Da No.70
IGDB-278	Improved	*G. max* (L.) Merr.	Jin Da No.62
IGDB-279	Improved	*G. max* (L.) Merr.	Jin Da No.52
IGDB-280	Improved	*G. max* (L.) Merr.	Jin Da No.26
IGDB-281	Improved	*G. max* (L.) Merr.	Ji Yu No.90
IGDB-282	Improved	*G. max* (L.) Merr.	Hei Nong No.51
IGDB-283	Improved	*G. max* (L.) Merr.	Hei He No.1
IGDB-284	Improved	*G. max* (L.) Merr.	He Feng No.25
IGDB-285	Improved	*G. max* (L.) Merr.	He Feng No.23
IGDB-286	Improved	*G. max* (L.) Merr.	He Dou No.13
IGDB-287	Improved	*G. max* (L.) Merr.	Gui Chun No.8
IGDB-288	Improved	*G. max* (L.) Merr.	Fen Dou No.89
IGDB-289	Improved	*G. max* (L.) Merr.	Fen Dou No.88
IGDB-290	Improved	*G. max* (L.) Merr.	Fen Dou No.86
IGDB-291	Improved	*G. max* (L.) Merr.	Fen Dou No.85
IGDB-292	Improved	*G. max* (L.) Merr.	Fen Dou No.79
IGDB-293	Improved	*G. max* (L.) Merr.	Fen Dou No.78
IGDB-294	Improved	*G. max* (L.) Merr.	Fen Dou No.65
IGDB-295	Improved	*G. max* (L.) Merr.	Fen Dou No.63
IGDB-296	Improved	*G. max* (L.) Merr.	Dong Nong No.52
IGDB-297	Improved	*G. max* (L.) Merr.	Dong Nong No.51
IGDB-298	Improved	*G. max* (L.) Merr.	Dong Nong No.26
IGDB-299	Improved	*G. max* (L.) Merr.	Chang Nong No.16
IGDB-300	Improved	*G. max* (L.) Merr.	Chang Nong No.15
IGDB-301	Improved	*G. max* (L.) Merr.	Chang Nong No.13
IGDB-302	Improved	*G. max* (L.) Merr.	Cang Dou-11

References

1. Doebley JF, Gaut BS, Smith BD: The molecular genetics of crop domestication. Cell 2006, 127:1309-1321.

2. Purugganan MD, Fuller DQ: The nature of selection during plant domestication. Nature 2009, 457:843-848.

3. Meyer RS, Purugganan MD: Evolution of crop species: genetics of domestication and diversification. Nat Rev Genet 2013, 14:840-852.

4. Olsen KM, Wendel JF: A bountiful harvest: genomic insights into crop domestication phenotypes. Annu Rev Plant Biol 2013, 64:47-70.

5. Li X, Scanlon MJ, Yu J: Evolutionary patterns of DNA base composition and correlation to polymorphisms in DNA repair systems. Nucleic Acids Res 2015, 43:3614-3625.

6. Harris K: Evidence for recent, population-specific evolution of the human mutation rate. Proc Natl Acad Sci U S A 2015, 112:3439-3444.

7. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A, et al: The simons genome diversity project: 300 genomes from 142 diverse populations. Nature 2016, 538:201-206.

8. Sharp PM, Matassi G: Codon usage and genome evolution. Curr Opin Genet Dev 1994, 4:851-860.

9. Bernardi G: The isochore organization of the human genome. Annu Rev Genet 1989, 23:637-661.

10. Bernardi G: Isochores and the evolutionary genomics of vertebrates. Gene 2000, 241:3-17.

11. Duret L, Galtier N: Biased gene conversion and the evolution of mammalian genomic landscapes. Annu Rev Genomics Hum Genet 2009, 10:285-311.

12. Springer NM, Schmitz RJ: Exploiting induced and natural epigenetic variation for crop improvement. Nat Rev Genet 2017, 18:563-575.

13. Song QX, Lu X, Li QT, Chen H, Hu XY, Ma B, Zhang WK, Chen SY, Zhang JS: Genome-wide analysis of DNA methylation in soybean. Mol Plant 2013, 6:1961-1974.

14. Glemin S, Clement Y, David J, Ressayre A: GC content evolution in coding regions of angiosperm genomes: a unifying hypothesis. Trends Genet 2014, 30:263-270.

15. Nachman MW: Variation in recombination rate across the genome: evidence and implications. Curr Opin Genet Dev 2002, 12:657-663.

16. Hershberg R, Petrov DA: Evidence that mutation is universally biased towards AT in bacteria. Plos Genet 2010, 6.

17. Mathieson I, Reich D: Differences in the rare variant spectrum among human populations. PLoS Genet 2017, 13:e1006581.

18. Massey DJ, Koren A: Mismatch repair prefers exons. Nature Genet 2017, 49:1673.

19. Hu Z, Cools T, De Veylder L: Mechanisms used by plants to cope with DNA damage. Annu Rev Plant Biol 2016, 67:439-462.

20. Ikehata H, Ono T: The mechanisms of UV mutagenesis. J Radiat Res 2011, 52:115-125.

21. Nawkar GM, Maibam P, Park JH, Sahi VP, Lee SY, Kang CH: UV-Induced cell death in plants. Int J Mol Sci 2013, 14:1608-1628.

22. Jones PA: Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet 2012, 13:484-492.

23. Law JA, Jacobsen SE: Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet 2010, 11:204-220.

24. Walser JC, Ponger L, Furano AV: CpG dinucleotides and the mutation rate of non-CpG DNA. Genome Res 2008, 18:1403-1414.

25. Tommasi S, Denissenko MF, Pfeifer GP: Sunlight induces pyrimidine dimers preferentially at 5-methylcytosine bases. Cancer Research 1997, 57:4727-4730.

26. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE: Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature 2008, 452:215-219.

27. Feng S, Jacobsen SE, Reik W: Epigenetic reprogramming in plant and animal development. Science 2010, 330:622-627.

28. West PT, Li Q, Ji L, Eichten SR, Song J, Vaughn MW, Schmitz RJ, Springer NM: Genomic distribution of H3K9me2 and DNA methylation in a maize genome. PLoS One 2014, 9:e105267.

29. Chia JM, Song C, Bradbury PJ, Costich D, de Leon N, Doebley J, Elshire RJ, Gaut B, Geller L, Glaubitz JC, et al: Maize HapMap2 identifies extant variation from a genome in flux. Nat Genet 2012, 44:803-807.

30. Zhou Z, Jiang Y, Wang Z, Gou Z, Lyu J, Li W, Yu Y, Shu L, Zhao Y, Ma Y, et al: Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat Biotechnol 2015, 33:408-414.

31. Li X, Zhu C, Yeh CT, Wu W, Takacs EM, Petsch KA, Tian F, Bai G, Buckler ES, Muehlbauer GJ, et al: Genic and nongenic contributions to natural variation of quantitative traits in maize. Genome Res 2012, 22:2436-2444.

32. Wallace JG, Bradbury PJ, Zhang N, Gibon Y, Stitt M, Buckler ES: Association mapping across numerous traits reveals patterns of functional variation in maize. PLoS Genet 2014, 10:e1004845.

33. Gore MA, Chia JM, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL, Peiffer JA, McMullen MD, Grills GS, Ross-Ibarra J, et al: A first-generation haplotype map of maize. Science 2009, 326:1115-1117.

34. Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, et al: Genome sequence of the palaeopolyploid soybean. Nature 2010, 463:178-183.

35. Wolfgruber TK, Sharma A, Schneider KL, Albert PS, Koo DH, Shi J, Gao Z, Han F, Lee H, Xu R, et al: Maize centromere structure and evolution: sequence analysis of centromeres 2 and 5 reveals dynamic Loci shaped primarily by retrotransposons. PLoS Genet 2009, 5:e1000743.

36. Lin JY, Jacobus BH, SanMiguel P, Walling JG, Yuan Y, Shoemaker RC, Young ND, Jackson SA: Pericentromeric regions of soybean (Glycine max L. Merr.) chromosomes consist of retroelements and tandemly repeated DNA and are structurally and evolutionarily labile. Genetics 2005, 170:1221-1230.

37. Wang Y, Tang X, Cheng Z, Mueller L, Giovannoni J, Tanksley SD: Euchromatin and pericentromeric heterochromatin: comparative composition in the tomato genome. Genetics 2006, 172:2529-2540.

38. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, et al: A high-resolution recombination map of the human genome. Nat Genet 2002, 31:241-247.

39. Rodgers-Melnick E, Bradbury PJ, Elshire RJ, Glaubitz JC, Acharya CB, Mitchell SE, Li C, Li Y, Buckler ES: Recombination in diverse maize is stable, predictable, and associated with genetic load. Proc Natl Acad Sci 2015:201413864.

40. Ikehata H, Ono T: Significance of CpG methylation for solar UV-induced mutagenesis and carcinogenesis in skin. Photochem Photobiol 2007, 83:196-204.

41. Wang P, Xia H, Zhang Y, Zhao S, Zhao C, Hou L, Li C, Li A, Ma C, Wang X: Genome-wide high-resolution mapping of DNA methylation identifies epigenetic variation across embryo and endosperm in Maize (Zea may). BMC Genomics 2015, 16:21.

42. Li Q, Gent JI, Zynda G, Song JW, Makarevitch I, Hirsch CD, Hirsch CN, Dawe RK, Madzima TF, McGinnis KM, et al: RNA-directed DNA methylation enforces boundaries between heterochromatin and euchromatin in the maize genome. Proc Natl Acad Sci U S A 2015, 112:14728-14733.

43. Kim KD, El Baidouri M, Abernathy B, Iwata-Otsubo A, Chavarro C, Gonzales M, Libault M, Grimwood J, Jackson SA: A comparative epigenomic analysis of polyploidy-derived genes in soybean and common bean. Plant Physiol 2015, 168:1433-1447.

44. El Baidouri M, Do Kim K, Abernathy B, Li Y-H, Qiu L-J, Jackson SA: Genic C-methylation in soybean is associated with gene paralogs relocated to transposable element-rich pericentromeres. Mol Plant 2018, 11:485-495.

45. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA: An integrated map of genetic variation from 1,092 human genomes. Nature 2012, 491:56-65.

46. Genomes Project C, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA: A map of human genome variation from population-scale sequencing. Nature 2010, 467:1061-1073.

47. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR: Deciphering signatures of mutational processes operative in human cancer. Cell Rep 2013, 3:246-259.

48. Ganpudi AL, Schroeder DF: UV damaged DNA repair & tolerance in plants. Intech Open Access Publisher; 2011.

49. Hufford MB, Xu X, van Heerwaarden J, Pyhajarvi T, Chia JM, Cartwright RA, Elshire RJ, Glaubitz JC, Guill KE, Kaeppler SM, et al: Comparative population genomics of maize domestication and improvement. Nat Genet 2012, 44:808-811.

50. Wang L, Beissinger TM, Lorant A, Ross-Ibarra C, Ross-Ibarra J, Hufford MB: The interplay of demography and selection during maize domestication and expansion. Genome Biol 2017, 18:215.

51. Beissinger TM, Wang L, Crosby K, Durvasula A, Hufford MB, Ross-Ibarra J: Recent demography drives changes in linked selection across the maize genome. Nat Plants 2016, 2:16084.

52. Zhou Y, Massonnet M, Sanjak JS, Cantu D, Gaut BS: Evolutionary genomics of grape (Vitis vinifera ssp. vinifera) domestication. Proc Natl Acad Sci U S A 2017, 114:11715-11720.

53. Marsden CD, Ortega-Del Vecchyo D, O'Brien DP, Taylor JF, Ramirez O, Vila C, Marques-Bonet T, Schnabel RD, Wayne RK, Lohmueller KE: Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs. Proc Natl Acad Sci U S A 2016, 113:152-157.

54. McCoy RC, Akey JM: Patterns of deleterious variation between human populations reveal an unbalanced load. Proc Natl Acad Sci U S A 2016, 113:809-811.

55. Liu Q, Zhou Y, Morrell PL, Gaut BS: Deleterious variants in Asian rice and the potential cost of domestication. Mol Bio Evol 2017, 34:908-924.

56. Zhang M, Zhou L, Bawa R, Suren H, Holliday JA: Recombination rate variation, hitchhiking, and demographic history shape deleterious load in poplar. Mol Biol Evol 2016, 33:2899-2910.

57. Akashi H, Gojobori T: Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. Proc Natl Acad Sci U S A 2002, 99:3695-3700.

58. Raiford DW, Heizer EM, Jr., Miller RV, Doom TE, Raymer ML, Krane DE: Metabolic and translational efficiency in microbial organisms. J Mol Evol 2012, 74:206-216.

59. Swire J: Selection on synthesis cost affects interprotein amino acid usage in all three domains of life. J Mol Evol 2007, 64:558-571.

60. Heizer EM, Raiford DW, Raymer ML, Doom TE, Miller RV, Krane DE: Amino acid cost and codon-usage biases in 6 prokaryotic genomes: A whole-genome analysis. Mol Bio Evol 2006, 23:1670-1680.

61. Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA: Selection for short introns in highly expressed genes. Nat Genet 2002, 31:415-418.

62. Li SW, Feng L, Niu DK: Selection for the miniaturization of highly expressed genes. Biochem Biophys Res Commun 2007, 360:586-592.

63. Chen WH, Lu G, Bork P, Hu S, Lercher MJ: Energy efficiency trade-offs drive nucleotide usage in transcribed regions. Nat Commun 2016, 7:11334.

64. Ussery DW, Wassenaar TM, Borini S: Computing for comparative microbial genomics: bioinformatics for microbiologists. Springer Science & Business Media; 2009.

65. Schuster-Bockler B, Lehner B: Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature 2012, 488:504-507.

66. Frigola J, Sabarinathan R, Mularoni L, Muinos F, Gonzalez-Perez A, Lopez-Bigas N: Reduced mutation rate in exons due to differential mismatch repair. Nat Genet 2017, 49:1684-1692.

67. Belfield EJ, Ding ZJ, Jamieson FJC, Visscher AM, Zheng SJ, Mithani A, Harberd NP: DNA mismatch repair preferentially protects genes from mutation. Genome Res 2018, 28:66-74.

68. Wicker T, Yu Y, Haberer G, Mayer KF, Marri PR, Rounsley S, Chen M, Zuccolo A, Panaud O, Wing RA: DNA transposon activity is associated with increased mutation rates in genes of rice and other grasses. Nat Commun 2016, 7:12790.

69. Muller HJ: Some genetic aspects of sex. Amer Nat 1932, 66:118-138.

70. Muller HJ: The relation of recombination to mutational advance. Mutat Res 1964, 106:2-9.

71. Felsenstein J: The evolutionary advantage of recombination. Genetics 1974, 78:737-756.

72. Charlesworth B: The evolution of sex chromosomes. Science 1991, 251:1030-1033.

73. Alexandrov LB, Stratton MR: Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. Curr Opin Genet Dev 2014, 24:52-60.

74. Eichten SR, Briskine R, Song J, Li Q, Swanson-Wagner R, Hermanson PJ, Waters AJ, Starr E, West PT, Tiffin P, et al: Epigenetic and genetic influences on DNA methylation variation in maize populations. Plant Cell 2013, 25:2783-2797.

75. Turunen M, Vogelmann T, Smith W: UV screening in lodgepole pine (Pinus contorta ssp. latifolia) cotyledons and needles. Int J Plant Sci 1999, 160:315-320.

76. Mazza CA, Boccalandro HE, Giordano CV, Battista D, Scopel AL, Ballaré CL: Functional significance and induction by solar radiation of ultraviolet-absorbing sunscreens in field-grown soybean crops. Plant Physiol 2000, 122:117-126.

77. Ries G, Heller W, Puchta H, Sandermann H, Seidlitz HK, Hohn B: Elevated UV-B radiation reduces genome stability in plants. Nature 2000, 406:98-101.

78. Meyerowitz EM: Plants compared to animals: the broadest comparative study of development. Science 2002, 295:1482-1485.

79. Mohrenweiser HW, Wilson DM, 3rd, Jones IM: Challenges and complexities in estimating both the functional impact and the disease risk associated with the extensive genetic variation in human DNA repair genes. Mutat Res 2003, 526:93-125.

80. Uchimura A, Higuchi M, Minakuchi Y, Ohno M, Toyoda A, Fujiyama A, Miura I, Wakana S, Nishino J, Yagi T: Germline mutation rates and the long-term phenotypic effects of mutation accumulation in wild-type laboratory mice and mutator mice. Genome Res 2015, 25:1125-1134.

81. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA, et al: Mutational landscape and significance across 12 major cancer types. Nature 2013, 502:333-339.

82. Hodgkinson A, Eyre-Walker A: Variation in the mutation rate across mammalian genomes. Nat Rev Genet 2011, 12:756-766.

83. Conrad DF, Keebler JE, DePristo MA, Lindsay SJ, Zhang Y, Casals F, Idaghdour Y, Hartl CL, Torroja C, Garimella KV, et al: Variation in genome-wide mutation rates within and between human families. Nat Genet 2011, 43:712-714.

84. Zhao H, Sun Z, Wang J, Huang H, Kocher J-P, Wang L: CrossMap: a versatile tool for coordinate conversion between genome assemblies. Bioinformatics 2013, 30:1006-1007.

85. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. J Mol Biol 1990, 215:403-410.

86. Young ND, Debelle F, Oldroyd GED, Geurts R, Cannon SB, Udvardi MK, Benedito VA, Mayer KFX, Gouzy J, Schoof H, et al: The Medicago genome provides insight into the evolution of rhizobial symbioses. Nature 2011, 480:520-524.

87. Li H: Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:13033997 2013.

88. Martin Morgan H, Maintainer MBP, ShortRead S, GenomicFeatures T, Biostrings L, biocViews DataImport I: Package 'Rsamtools'. 2013.

89. Durinck S, Bullard J, Spellman PT, Dudoit S: GenomeGraphs: integrated genomic data visualization with R. BMC Bioinformatics 2009, 10:2.

90. Scheet P, Stephens M: A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 2006, 78:629-644.

91. Paradis E: pegas: an R package for population genetics with an integrated-modular approach. Bioinformatics 2010, 26:419-420.

92. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM: A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 2012, 6:80-92.

93. Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, Campbell MS, Stein JC, Wei X, Chin C-S: Improved maize reference genome with single-molecule technologies. Nature 2017, 546:524.

94. Du J, Grant D, Tian Z, Nelson RT, Zhu L, Shoemaker RC, Ma J: SoyTEdb: a comprehensive database of transposable elements in the soybean genome. BMC Genomics 2010, 11:113.

95. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, et al: A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 2006, 38:203-208.

96. Zhang Z, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordovas JM, Buckler ES: Mixed linear model approach adapted for genome-wide association studies. Nat Genet 2010, 42:355-360.

97. Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, Gore MA, Buckler ES, Zhang Z: GAPIT: genome association and prediction integrated tool. Bioinformatics 2012, 28:2397-2399.

98. Wang J, Li X, Kim KD, Scanlon MJ, Jackson SA, Springer NM, Yu J: Genome-wide nucleotide patterns and potential mechanisms of genome divergence following domestication in maize and soybean source code. GitHub 2019, https://doi.org/10.5281/zenodo.2566552.

CHAPTER 3. AERIAL BASED HIGH-THROUGHPUT PHENOTYPING FOR GENETIC DISSECTION OF NDVI IN MAIZE

Jinyu Wang[1], Xianran Li[1], Matthew Dzievit[1], Xiaoqing Yu[1], Kevin P. Price[2], Jianming Yu[1]*

[1]Department of Agronomy, Iowa State University, Ames, IA 50011, USA
[2]Air Data Solutions, George West, TX 78022, USA
*Correspondence should be addressed to J.Y. (jmyu@iastate.edu)

Modified from a manuscript to be submitted to Plant Physiology

Abstract

Plant phenotyping under field conditions plays an important role in agricultural research. Efficient and accurate high-throughput phenotyping strategies enable a better connection between genotype and phenotype, which is critical for crop improvement. Unmanned aerial vehicle-based high-throughput phenotyping platforms (UAV-HTPPs) provide novel opportunities for large-scale proximal measurement of plant traits with high efficiency, high resolution, and low cost. In this study, we extracted time series NDVI data from multispectral images at 5 time points across the growing season of 1,752 diverse maize accessions with a UAV-HTPP. We identified genotypic differences and analyzed the dynamics and developmental trends of NDVI during different maize growth stages. Clustering analysis with time series NDVI classified 1,752 maize accessions into 2 groups possessing distinct NDVI developmental trends. Then the time series NDVI data were used in penalized-splines (P-splines) model to obtain genotype-specific curve parameters. Genome-wide association study (GWAS) using static NDVI values observed from individual time points and P-splines estimated NDVI curve parameters as phenotypic traits detected signals significantly associated with the traits. Additionally, GWAS for P-splines fitted NDVI values discovered the dynamic change of SNP effect for the trait associated genetic loci, which may suggest the role of gene-environment interplay in controlling NDVI development. Our results suggest the usefulness of UAV-based remote sensing for genetic dissection of NDVI.

Introduction

Crop production per unit area has to be doubled by 2050 to meet the future demand of food and fiber from the increasing global population (Gerland et al., 2014). Although agricultural research has improved the crop yields dramatically over the past few decades, the rates of genetic improvement for many crops are still below what is needed to meet the future demand (Ray et al., 2013). A major challenge for crop improvement is to establish the connection between phenotype and its genotype (White et al., 2012). The advances of sequencing and genotyping technologies over the past decade have improved the genotyping efficiency and provided a huge amount of genomic data, but the transition of these data into the identification of desirable traits is constrained by the ability of efficient phenotyping (White et al., 2012; Cobb et al., 2013). Phenotyping under field conditions has become the bottleneck for crop improvement (Cobb et al., 2013).

In recent years, there has been increased interest in field-based, high-throughput phenotyping platforms (HTPPs) using ground wheeled or aerial vehicles with multiple types of sensors, particularly for applications in breeding and germplasm evaluation (Furbank and Tester, 2011; Fiorani and Schurr, 2013; Walter et al., 2015). Ground-based phenotyping platforms have significantly improved the phenotyping efficiency (Andrade-Sanchez et al., 2014; Fernandez et al., 2017), but they do have limitations in terms of the scale it can be used, the portability, the ability to measure different crop systems, and the time required to measure large number of plots at different field locations (Haghighattalab et al., 2016). Aerial-based phenotyping platform is a great complement to the ground-based platform as it enables the rapid characterization of many plots and large-scale crop condition monitoring due to the high spatial and spectral resolutions of the sensors (Chapman et al., 2014; Haghighattalab et al., 2016).

One of the emerging technologies in aerial-based phenotyping platforms is unmanned aerial vehicle-based HTPPs (UAV-HTPPs), which are powerful remote sensor-bearing platform and are generally thought of as a cost-effective tool for crop phenotyping and precision agriculture (Hunt et al., 2005; Berni et al., 2009; Dunford et al., 2009; Chao et al., 2010; Zhang and Kovacs, 2012; Ballesteros et al., 2014; Chapman et al., 2014; Liebisch et al., 2015). UAV-HTPPs are able to assess a large number of plots almost simultaneously to minimize the effect of varied environmental conditions such as cloud cover, wind speed, and solar radiation (Chapman et al., 2014; Haghighattalab et al., 2016). Remote sensing phenotyping techniques are mainly based on information provided by visible/near-infrared radiation reflected and far-infrared emitted by plants (Berger et al., 2010; Vadivambal and Jayas, 2011; Zia et al., 2013). Remote sensing is non-destructive and resource conservative, which allows repeat inventory and measurement of physiological characteristics and detects change over time (Rundquist et al., 2001). When collected from UAV-HTPPs, remote sensing allows synoptic visualization, mapping, assessment and quantification of physiological characteristics of vegetation like biomass and relative stress or vigor (Yang et al., 2017).

Normalized difference vegetation index (NDVI), an effective leaf greenness indicator, is the most popular trait summarized from remote sensing. NDVI is derived from the difference between the reflectance in the visible red spectral region (λ = 500-700 nm) and the reflectance in the near infrared region (NIR, λ = 760-900 nm) (Kumar and Silva, 1973). Healthy plants generally have low reflectance in the red spectral region as a result of chlorophyll absorption and high reflectance in the NIR spectral region as a result of leaf cellular structure and canopy architecture. Therefore, NDVI can predict plant photosynthetic activity which is determined by the chlorophyll content and activity. NDVI is known to be associated with many traits, such as leaf chlorophyll content (SPAD), leaf area index (LAI), stay-green and senescence, nitrogen

usage efficiency, drought-adaptive traits, biomass, and grain yield (Duncan et al., 1967; Bort et al., 2005; Liebisch et al., 2015; Condorelli et al., 2018).

Remote sensing with UAV-HTPPs has been successfully adopted to measure NDVI in many experiments. But many studies have been limited to a single date and only a few studies have reported the use of UAV-HTPPs remote sensing to analyze the dynamics and developmental trends of NDVI among a large number of genotypes. The dynamics of plant growth captures critical biological information. Investigating the dynamics of plant growth may assist the detection of genotypic differences that can not be detected by single time point data. Our study reports the use of UAV-based NDVI remote sensing for genome-wide association study (GWAS) analysis in maize. This work presents a case of how a UAV-HTPP can be used for field maize NDVI measurement of a large number of genotypes at multiple growth stages across the growing season and genetic dissection of NDVI acquired from UAV-HTPPs. Here we report findings from the analysis and evaluation of genotypic differences and dynamic changes with time series NDVI data in 1752 maize accessions. We also performed the genetic dissection of NDVI with GWAS using both NDVI of individual time points and curve parameters developed from time series NDVI.

Materials and Methods

Field experiment

The study was conducted at Ag Engineering and Agronomy Research Farm at Boon, IA, US (42°01'10.26774"N, 93°46'11.48730"W). The experiment consists of 1752 diverse maize inbred lines, referred as 1752 Ames Panel accessions, sampled from USDA-ARS NCRPIS collection (Romay et al., 2013). Based on the classification from the original study (Romay et al., 2013) and the observed kernel structure, there are 151 non-stiff stalk lines, 108 popcorn lines, 149 stiff stalk lines, 134 sweet corn lines, 47 tropical lines, and 1163 unclassified lines. These

1752 accessions were planted on May 8, 2017 at ~30,000 plants per acre in 3.81 m long plots and spaced 0.76 m apart in an augmented randomized complete block design. Flowering time, plant height, and ear height were measured for each accession on a plot basis. Flowering time was recorded as the number of days after planting when 50% or more of the plants in a plot were shedding pollen. Plant height and ear height were measured from the representative plant in a plot as the distance between ground to flag leaf and the distance between ground to ear, respectively.

Unmanned aerial vehicle system

The UAV system contains a DJI S900 UAV and an NIR converted multispectral Canon Rebel SL1 DSLR camera with an intervalometer and GPS (Figure 1A). The DJI S900 is a highly wind resistant hexacopter and can be flown at different altitudes to collect data at the sub-cm resolution. The multispectral Canon camera is modified to sense the visible green and red regions as well as the invisible near infrared (NIR) regions of the electromagnetic spectrum (Green, 500-620 nm; Red, 550-720 nm; NIR, 800-900 nm).

Remote sensing data collection

We conducted 5 UAV overflights across the growing season in 2017. Overflights were scheduled around 5 growth stages (V_4, V_8, V_{12}, V_T, and R_5) as shown in Figure 1B. On the ground, we installed 12 white crosses with ~76 cm stripes that are evenly distributed across the field as ground control points (GCPs). The coordinates of the GCPs were measured by Real-Time Kinematics (RTK) GPS (Ashtech GPS/GNSS surveying systems – MODEL PROMARK 200) and used to georeference orthomosaics in the image processing. ISO sensitivity was 400 and shutter speed was 1/800 s for the multispectral camera. Under the control of an autopilot system with GPS, the UAV flew along pre-defined flight routes designed by the PC Ground Station software (DJI Co., Ltd., Shenzhen, China). The flight routes were designed to have at

least 80% overlap of images. Images were taken at an altitude of ~37 m and ~600 images of ~3 cm resolution were taken for each overflight.

Image processing

Following image processing steps were applied to obtain high quality data from the raw UAV images (Figure 1C-E): image pre-processing, orthomosaic generation, vegetation indices (VIs) calculation, and plot-level data extraction. Our camera was calibrated with Lambertian calibration panel at the beginning of each flight to compensate for differences in incoming solar radiation, which influence reflectance. Hundreds to thousands of raw images captured flying over each field were pre-processed for lens distortion, chromatic aberration, and gamma correction using ArcGIS Spatial Analyst Tools. Then the pre-processed images were used to generate the orthomosaic of the field. Orthomosaic generation comprised five main steps namely loading and aligning photos, importing ground control point (GCP) positions and georeferencing, building dense point cloud, building digital elevation model (DEM), and finally generating the orthomosaic image. Orthomosaiking allows for proper planimetric positioning of image pixels by reducing image distortion and correction for the bidirectional reflectance distribution function phenomena that are characteristic of images acquired from multiple viewing angles relative to incoming solar radiation angles. With the orthomosaic, a map showing the vegetation indices (VIs) were generated. To extract plot level data, we defined individual plot boundaries from orthomosaic image with an assigned plot ID that identifies the genotype and laid transects with 30 cm buffer-size for each plot. Then plot-level NDVI mean was calculated from the reflectance measurements in the red and NIR portion of the spectrum from the transect area of each plot. The equation for NDVI calculation as:

$$NDVI = (R_{NIR} - R_{red})/(R_{NIR} + R_{red})$$

where R_{NIR} is the reflectance measurements in the NIR spectrum, and R_{red} is the reflectance measurements in the red band spectrum. Values of NDVI range from 0 to 255. To avoid the effect of soil, a segmentation process based on the NDVI excluded non-plant material and obtained the NDVI values of the area covered by the plant ($NDVI_{plant}$). $NDVI_{plant}$ is used for the analysis in this study.

Sequence information and SNP extraction

We used two sets of SNPs, the raw SNP set and the imputed SNP set, for the analysis of 1752 Ames Panel accessions in this study. Both raw and imputed SNP sets use the B73 genome (AGPv3) as references.

The raw SNP set was extracted from genotypes of ZeaGBSv2.7, which is available at /iplant/home/shared/panzea/genotypes/GBS/v27/ZeaGBSv27_publicSamples_rawGenos_AGPv3_20170206.vcf.gz. To obtain the raw SNP set, we first combined the SNP call for accessions with multiple GBS samples in ZeaGBSv2.7 to get a consensus SNP set for the 2812 Ames Panel accessions. Then we extracted the SNP for 1752 accessions from the consensus SNP set. We obtained the raw SNP set with 316,047 SNPs by further filtering with a minor allele frequency (MAF) threshold of 1% and a missing rate threshold of 20%. The imputed SNP set for the 1752 Ames Panel accessions was projected from the 282 maize association panel which has high SNP density. After imputation, we obtained 16,792,800 SNPs by extracting SNPs for the 1752 Ames Panel accessions from the newly imputed SNP set and further filtering with a MAF threshold of 1%.

Clustering and population structure analysis

Average silhouette method (Tibshirani et al., 2001) and gap statistic (Kaufman and Rousseeuw, 1990) in R package *factoextra* were used to determine the optimum number of clusters for NDVI values at 5 time points of the 1752 maize accessions. To further validate the

clustering result from *K*-means, we conducted t-Distributed Stochastic Neighbor Embedding (t-SNE) (Van Der Maaten and Hinton, 2008) on NDVI values at 5 time points with R package *Rtsne.* T-SNE was also carried out on the raw SNP set of the 1752 maize accessions to study if the grouping pattern from SNP and NDVI agrees with each other. Principal component analysis (PCA) with the raw SNP set containing 316,047 SNPs was carried out with Genome Association and Prediction Integrated Tool (GAPIT) (Lipka et al., 2012).

Statistical modeling of NDVI growth curve

To summarize the general trend captured by the NDVI across the growing season, we fitted the penalized-splines (P-splines) model (Eilers and Marx, 1996) with the NDVI of 5 surveyed time points using *PsplinesREML* function of R package *splines*. We then obtained 3 curve parameters that are asymptote, max rate, and inflection point, for each accession. With the established P-splines model for each accession, we inferred the NDVI value for each accession at every one-week window between the start overflight date 37 DAP and the end overflight date 115 DAP (44, 51, 58, 65, 72, 79, 86, 93, 100, 107, 114 DAP) for further analysis.

Pearson correlation test between the observed NDVI and P-splines model fitted NDVI for each of the 5 time points were performed to check how well the P-splines can model the time series NDVI.

GWAS for NDVI

A mixed linear model (MLM) with both fixed covariates and a random kinship matrix (Yu et al., 2006; Zhang et al., 2010) was used to detect SNPs associated with the traits under study in GAPIT version 3.35 (Lipka et al., 2012). Parameters in MLM were determined by model selection process. As our genotype data was imputed, linkage disequilibrium is expected to exist between SNPs. The bonferroni method of multiple testing adjustment is easy to compute, but is well known to be conservative in the presence of LD. SimpleM (Gao et al., 2008, 2010;

Gao, 2011) is an efficient and accurate method for multiple testing adjustment when there is high LD in the SNP data set. Thus, we used simpleM to obtain the effective number of individual tests (M_{eff}). PCA-cutoff in simpleM was 0.995. With simpleM, the M_{eff} is 4,407,833. Then we used this ~4.4M M_{eff} for the Bonferroni correction to get the significance threshold.

Results

Dynamics of NDVI values across the growing season

We first examined the distribution of NDVI for the 5 surveyed time points. The average NDVI values from 1,752 accessions increased at a very fast rate during the time interval from 37 DAP to 44 DAP (NDVI from 115.39 to 123.64) and kept increasing but with a gradually decreasing rate after that time interval (Table 1, Figure 2). There is a negligible increase in NDVI mean values from 73 DAP to 115DAP (NDVI from 129.08 to 129.44). Different time points show different degrees of variations in NDVI. Variations of NDVI at the first and last time points are large (coefficient of variation (CV) ≥ 2.7), and relatively small for the 3 time points in between (CV ≤ 1.96). This may suggest that NDVI at the early and late time points can better differentiate genotypes than time points in the middle.

We compared the correlations between NDVI at different time points. As shown in Figure S1, NDVI from adjacent time points have high positive correlations with each other, and the magnitude of correlation decreases when two time points are far apart from each other. For example, the Pearson correlation coefficient between NDVI on 37 DAP and 44 DAP is 0.78, but correlations between NDVI on 37 DAP and that on 60 DAP and 73 DAP are 0.48 and 0.28, respectively.

NDVI is known to be positively correlated with chlorophyll content and green leaf area (Wu et al., 2008). Plant senescence is usually associated with chlorophyll degradation (Schippers et al., 2015) and flowering date of a plant may indicate how quickly it transfers from the

vegetative stage to the reproductive stage and may provide information on how soon plant will enter the senescence stage. Flowering date is reported to be correlated with plant height and ear height in maize (Troyer and Larkins, 1985). We then evaluated the correlation between NDVI from 5 surveyed time points and manually measured flowering time, plant height, and ear height. As shown in Figure S2, NDVI at the first time point is weakly and negatively correlated with flowering time. With the progress of the growing season, the correlation between flowering time and NDVI becomes positive and the magnitude of correlation keeps increasing, reaching maximum at the last time point ($r = 0.553$). This suggests that early flowering accessions generally senescence earlier, consequently having smaller NDVI values at late time points. This is also supported by the fact that on 115 DAP, the last surveyed time point, early flowering sweet corn group has a small NDVI mean value, while the late flowering tropical group has a larger NDVI mean value (Table 1). NDVI across all the surveyed time points are weakly and positively correlated with plant height, and the correlations between them reach maximum on 73 DAP. The same kind of correlation relationship exists between NDVI and ear height (Figure S2).

Different NDVI profile for sweet corn

When comparing the NDVI distribution for different groups, it seems sweet corn group behaves differently compared with the remaining 5 groups: *a*) sweet corn generally has smaller NDVI values compared with the remaining 5 groups; *b*) while NDVI values of all the other groups keep increasing across the growing season, NDVI values of sweet corn decrease towards the end of the growing season.

Using clustering analysis, genotypes with a similar trait change trend across the growing season will be grouped together. Thus, clustering analysis can help us distinguish the dynamic change of the target trait in the time dimension. To better understand the dynamic change of NDVI across the growing season for 1752 maize accessions, we conducted clustering analysis on

NDVI values of the 5 time points. Two well-known methods, average silhouette (Tibshirani et al., 2001) and gap statistic (Kaufman and Rousseeuw, 1990), were used to determine the optimum number of clusters. As shown in Figure S3, both silhouette and gap statistics methods indicate that the optimum number of clusters should be 2. We then clustered the 1752 accessions into 2 clusters.

With *K*-means, 572 accessions are classified into one cluster and 1180 accessions are classified into the other cluster, referred as cluster 1 and cluster 2, respectively (Table 2 and Figure S4). Excluding accessions belonging to the unclassified group, cluster 1 composed of mostly sweet corn accessions (51.9%) and cluster 2 composed mostly of accessions from stiff stalk (29.7%), non-stiff stalk (28.5%), and popcorn (20.8%). We then visualized the clustering result by plotting out the NDVI growth curves (Figure 3A-B). NDVI values of cluster 2 accessions either kept increasing across all the surveyed time points or reached and stayed at the plateaus at late time points, while that of cluster 1 increased during the first 2 time points, reached the plateaus around the 3rd or 4th time point, and decreased after reaching the plateaus. In addition, cluster 2 consistently has higher average NDVI values than cluster 1 across all the 5 surveyed time points.

Then we conducted t-SNE (Van Der Maaten and Hinton, 2008), a dimension reduction method, with NDVI values of the 5 growth stages to validate the *K-means* clustering result. Accessions classified to cluster 1 and cluster 2 by *K-means* are clearly separated from each other with t-SNE, which suggests the grouping pattern from t-SNE agrees with the *K-means* clustering result (Figure 3C-D).

NDVI dynamics revealed by the clustering analysis encouraged us to study if the growth pattern can be explained by the genotype information. We then conducted t-SNE and PCA with 316,047 SNPs from the raw SNP set. Excluding the unclassified group, the rest 5 groups in

general can be separated by both t-SNE and PCA with SNP information (Figure 3F, Figure 3H). For example, sweet corn is located on one side of the 1st dimension of t-SNE, while popcorn on the other side. Incorporating grouping patterns from genotype and time series NDVI, we can see that the observed NDVI dynamics can be partially explained by genotypic information. In general, cluster 1 and cluster 2 from *K-means* clustering on NDVI can be well separated by t-SNE and PCA using SNP information (Figure 3E, Figure 3G). The majority accessions belonging to cluster 1 located at the left side along principal component 1 (PC1) axis and 1st dimension of t-SNE. And accessions belonging to cluster 2 mainly located at the right side along these two axes.

Statistical modeling of time series NDVI

Plant physiological traits measured across the growing season contain rich biological information. P-splines has been proved to be effective in quantitatively summarizing such data and to be flexible in modeling different shapes of developmental curves (Calderon et al., 2010; Hurtado et al., 2012). To further explore the biological information embedded in time series NDVI of 1752 maize accessions, we modeled the NDVI values of the 5 surveyed time points with P-splines.

Using P-splines, we were able to fit NDVI curves for 1751 accessions. P-splines fitted NDVI curves are very similar to the curves developed from observed NDVI (Figure 3A, Figure 4A). In general, similar to what we observed earlier, two different growth patterns were revealed from the P-splines fitted NDVI curves. One growth pattern is that NDVI values increased very fast at early time points, reached the plateaus at the middle time points, and decreased at late time point, and the other growth pattern is that NDVI values either kept increasing across all 5 time points or reached and stayed at the plateaus towards the late time points. Growth rates for the majority of the maize accessions decrease from early to late time points (Figure 4B). To

assess how well P-splines fit the time series NDVI, we calculated Pearson correlation coefficients between the observed NDVI values and model fitted NDVI values for each of the 5 surveyed time points. Overall, there are strong correlations between the observed and P-splines fitted NDVIs (*r* ranging from 0.881 – 0.999) (Figure 4C-D, Figure S5), with relatively low correlation at the early time point and relatively high correlation at late time point. This may indicate that P-splines can model time series NDVI better at late growth stages than early growth stages.

After fitting the curves for each accession, 3 curve parameters - asymptote, max rate, and inflection point that capture different features of NDVI curves were estimated. Asymptote is the model fitted maximum NDVI value. Max rate is the model estimated maximum growth rate of NDVI. Inflection point is the point in time with maximum growth rate of NDVI. Asymptote and max rate follow a normal distribution. Inflection point is mostly constant for 1751 maize accessions. Thus, asymptote and max rate were used as phenotypic traits in the following GWAS analysis but not inflection point. We then studied the relationship between observed NDVI values with asymptote and max rate (Figure S6). Asymptote is correlated with observed NDVI, with strong correlation between them at late time points and weak correlation at early time points. Max rate is weakly correlated with observed NDVI across all the surveyed time points.

GWAS for NDVI of individual time points and curve parameters

With genome-wide association studies (GWAS), the previous study in durum wheat detected quantitative trait loci (QTLs) for NDVI (Condorelli et al., 2018). While several studies have reported the analysis of NDVI obtained from aerial-based sensors in maize (Liebisch et al., 2015; Zaman-Allah et al., 2015; Han et al., 2018; Yonah et al., 2018), to our knowledge, no specific studies have so far explored the usefulness of UAV-based NDVI measurements for genetic dissection of NDVI in maize. In this study, we conducted genome scans with 16,792,800

SNPs in the imputed SNP set for observed NDVI from individual time points, P-splines estimated asymptote and max rate, and P-splines fitted NDVI to dissect the genetic structure underlying NDVI. As there are correlations between NDVI and manually measured flowering time, plant height, and ear height (Figure S2), genome scans for these 3 manually measured traits were also conducted to see if these traits have shared signals with NDVI.

Using observed NDVI values of individual time points as phenotypes, genome scans identified 4 strong associations (Figure 5, Figure S6). Among these 4 associations, two of them were detected by NDVI on 37 DAP, one was detected by NDVI on 44 DAP, and one was detected by NDVI on 115 DAP. Both associations detected by NDVI on 37 DAP are located on chromosome 8, with one near the start of the chromosome and the other close to the end. The association close to the end of the chromosome lies 0.2 Mb downstream of gene *GRMZM2G316907*, a putative orthologue of the *Arabidopsis AT3G47570.1,* encoding leucine-rich repeat protein kinase family protein which functions in controlling cell proliferation and meristem maintenance (Torii, 2004). For the association near the start of chromosome 8, no candidate gene was identified. The best association detected by NDVI on 115 DAP is located on chromosome 8, and it lies 2 kb upstream of gene *GRMZM2G094241,* a putative orthologue of the *Arabidopsis KNAT6* (knotted1-like homeobox gene 6) which is expressed in vegetative meristem and functions in controlling leaf morphology (Lincoln et al., 1994).

Genome scans with P-splines estimated NDVI curve parameters identified a number of strong associations (Figure 5). Max rate detected a number of strong associations, but asymptote did not detect any significant associations. 4 potential candidate genes were identified by surveying annotated maize genes located within 100 kb to the associated signal regions of max rate (Table. S2). For example, the second best hit of max rate on chromosome 2 is located 7 kb upstream of gene *GRMZM2G002043*, a putative orthologue of the *Arabidospsis PDM2*

(Pentratricopeptide Repeat Protein Pigment-Defective Mutant2) that functions regulating plastid gene expression required for normal chloroplast development (Du et al., 2017). And the second best hit on chromosome 3 is located within gene *GRMZM2G114399*, a putative orthologue of the *Arabidospsis PPD5* (Mog1/PsbP/DUF1795-like photosystem II reaction center PsbP family protein) that functions in the photosynthetic pathway (Roose et al., 2011). While asymptote did not detect any loci that pass the significance threshold, its strongest signal on chromosome 8 detected the same peak region as the strongest signal from NDVI on 115 DAP.

We obtained 15 P-splines model fitted NDVI values for each accession that corresponding to each of the 5 UAV overflight dates (37, 44, 60, 73, and 115 DAP) and each one-week window between the start and end overflight date (51, 58, 65, 72, 79, 86, 93, 100, 107, 114 DAP). Genome scan for these 15 P-splines fitted NDVI detected two significant associations. One of the associations is located on chromosome 8 and detected by observed NDVI on 115 DAP. This association signal starts to show up for the fitted NDVI on 72 DAP, becomes significant on 79 DAP and gradually become more significant after this time point, and reaches the maximum significance level on 115 DAP. The other association is located on chromosome 7 and was not detected by any of the 5 observed NDVI but picked up by P-splines fitted NDVI values on 72 DAP and 73 DAP. This association signal lies 1 kb away of gene *GRMZM2G300709,* a putative orthologue of *Arabidopsis POK1* (phragmoplast orienting kinesin 1) that functions in establishing the cortical division (Lipka et al., 2014). The association signal for this gene starts to show up on 65 DAP, becomes significant on 72 and 73 DAP, and changes to be non-significant after 79 DAP. The change of significance level across the growing season for genes identified by P-splines fitted NDVI encouraged us to study the dynamic change of the SNP effect. Notably, the additive effect size and direction of the most significant SNPs for those two genes are time/environment-dependent (Figure 6). Alleles increasing NDVI in one time

point can increase NDVI to a different level, decrease NDVI, or have no effect in other time points. This may suggest that gene-environment interplay plays an important role in controlling the development of NDVI.

The genome scan for flowering time detected two known flowering time genes *ZCN8* (*GRMZM2G179264*) and *VGT1* (*GRMZM2G700665*) on chromosome 8 (Salvi et al., 2007; Romero Navarro et al., 2017) (Figure S8). The genome scan for plant height detected the known plant height genes *Br2* (Multani et al., 2003; Xing et al., 2015) and a few genes function in the senescence pathway. The signal for flowering time gene *ZCN8* was also detected by plant height and ear height. None of the significant loci were shared between NDVI and 3 manually measured traits, which probably due to the low correlation between them (Figure S2).

Discussion

NDVI variation

The CV of NDVI at the first time point was relatively large and decreased at the second and third time points. From the fourth time point, the CV of NDVI started to increase and reached the maximum level at the last time point. The large NDVI variation at the first time point was likely affected by the different emergence time of different genotypes. At a very early growth stage, accessions emerged early had more leaves, larger green leaf area, and larger NDVI values than accessions emerged late. And the large NDVI variation at the last time point could be due to different senescence status of different genotypes. At the late growth stage, early senescence genotypes had more yellow leaves, smaller green leaf area, and smaller NDVI values than late senescence genotypes. Across the whole growing season, the trend of NDVI variation in this diverse maize population is expected to be: 0 NDVI variation right after planting as no genotypes have emerged; then the NDVI variation increases and reaches maximum because different emergence time of different genotypes; with the progress of growing season, it reduces

and stays at a similar level as all the genotypes enter similar developmental stages; then NDVI variation increases again because different senescence status of different genotypes.

The increase in CV during the late time points may suggest that genotypic differences start to show up when plants gradually enter the reproductive stages and accumulate after that time point. The changed variation patterns of NDVI across the growing season indicates that different growth stages have differed capacities to discriminate genotypes. Previous studies have shown that NDVI remote sensing platforms differ in their capacity to discriminate genotypes, especially depending on the plant developmental stage (Marti et al., 2007; Christopher et al., 2016). Whether the different NDVI variation patterns at different time points observed in this study is due to the remote sensing platforms or not cannot be verified.

Correlation between NDVI and manually measured traits

We observed low to medium strength correlation between flowering time and NDVI for the 5 surveyed time points, with weak negative correlation at the early time points and medium strength positive correlation at the last time point. Together, this may suggest that flowering time of plants have increased effect on NDVI at late growth stages across the growing season.

The correlation between plant height and NDVI of all 5 time points is generally low ($r <$ 0.4), which is similar to what has been observed in previous study (Han et al., 2018). Compared with the previous study, the slightly lower correlation between these two traits on 60 DAP and slightly higher correlation between them on 73 DAP in our study could be because the plant height and NDVI are not measured at the same growth stage. The plant height in this study was only measured once at the end of the growing season.

Clustering analysis revealed NDVI dynamics

Clustering analysis of time series trait data is able to group genotypes with similar trait dynamics together. Thus, clustering analysis can help discriminate genotypic differences in time

dimension. Here, the 1752 maize accessions were classified into 2 clusters that have distinct NDVI change patterns using the *K-means* clustering method, which agrees with the NDVI change pattern at each of the 5 known groups. Among the 5 groups, stiff stalk, non-stiff stalk, popcorn, and tropical groups have similar NDVI dynamics across the 5 surveyed time points. NDVI values of these 4 groups keep increasing across the growing season. Tropical group has a fast increase in NDVI value from the 4^{th} to 5^{th} time point, which might be due to tropical accessions' rejuvenating as a result of the favorable environment condition after flowering. Sweet corn group exhibit a different NDVI dynamics than the other 4 groups. NDVI values of sweet corn accessions have an obvious decreasing trend towards the end of the growing season. One possible reason for the distinct NDVI growth curve of the sweet corn group is that sweet corn accessions were derived from Northern Flint materials and were adapted for temperate climate (Romay et al., 2013). We also tried to classify the 1752 maize accessions into 3 and 4 clusters. However, when the number of clusters is larger than 2, the NDVI dynamics of different clusters are not clearly distinguishable from each other.

NDVI measurements by UAV-HTPPs

In this study, we obtained UAV-based NDVI around 5 maize growth stages across the growing season. Taking all the 1752 maize accessions together, it looks like NDVI values kept increasing across all 5 surveyed time points (from 37 to 115 DAP), which seems to be different as previously observed NDVI curves (Govaerts and Verhulst, 2010; Wang et al., 2016; Han et al., 2018). Previously observed NDVI curves are similar to the bell shape. NDVI values in general keep increasing during the early growth stage, reach the maximum at the tasseling stage and decrease after that. We suspect the main reason for this difference is that our UAV overflights did not catch the tasseling stage for the majority of the 1752 maize accessions. As shown in Table 1, the mean flowering time of 1752 maize accessions is ~78 DAP and the

maximum flowering time is 118 DAP, while the closest UAV overflight date to tasseling stage is 73 DAP. It is likely that the missing NDVI measurement around tasseling stage generated the non-decreasing NDVI curve. This point is supported by the bell shape like NDVI curve of the sweet corn group. The average flowering time for the sweet corn group is ~71 DAP which is only two days away from our fourth UAV overflight on 73 DAP. Although we scheduled to conduct the UAV overflight at the maize tasseling stage, we were not able to catch this growth stage due to the unfavorable weather condition for UAV overflight during this time. In future experiments, we should plan to survey more growth stages and have more frequent UAV overflights to resolve this potential issue.

GWAS analysis of NDVI and curve parameters

This study presents one case for the use of UAV-based NDVI remote sensing for GWAS analysis in a large maize population. We conducted GWAS with NDVI of individual time points. We also modeled time series NDVI first with P-splines to obtain genotypic-specific curve parameters, and then use these curve parameters as phenotypic traits in conventional GWAS analysis. Although time series NDVI that capture developmental information were obtained, genetic analysis of NDVI at individual time points are not able to study NDVI as a continuous trait. Parameters estimated from NDVI growth curves are able to summarize and incorporate information from all time points, which capture rich biological information and enable us to study it as a developmental process. Combining information over time may also help reduce measurement errors from the UAV-HTTP platforms.

We obtained 3 curve parameters, asymptote, max rate, and inflection point by modeling time series NDVI with P-splines. But we only used asymptote and max rate for the GWAS analysis. We decided not to use the inflection point for GWAS as it is mostly constant. We think that the constant value of the inflection point is because the low UAV overflight frequency in the

early growing season is not able to collect enough data points to allow the P-splines model to generate different values for this parameter. In future experiments, we should increase the overflight frequency at the fast-growing stage to have enough data points to generate different values for the inflection point. We could also extrapolate the NDVI growth curve outside of the growth window covered by the 5 UAV overflights to help sample time point for the future UAV overflight, but this needs to be done with caution since P-splines is mainly for data fitting.

GWAS with P-splines fitted NDVI values also discovered the dynamic change of the SNP effect for trait associated genetic loci, which may suggest the important role of gene-environment interplay in controlling NDVI development. *KNAT6* functions in shoot apical meristem (SAM) maintenance (Lincoln et al., 1994). *POK1* plays a role in division plane maintenance at the cell cortex (Lipka et al., 2014). Both *KNAT6* and *POK1* can affect leaf morphology. One possible explanation for the dynamic change of the SNP effect for gene *KNAT6* and *POK1* is that there are two groups of accessions having different sequence polymorphisms for these two genes, and these polymorphisms may affect the difference in gene expression of these two groups. The magnitude of difference in gene expression level between these two groups is further affected by environment factors or developmental stages.

Except for the shared GWAS signals between NDVI on 115 DAP and asymptote on chromosome 8, we didn't find other strongly shared GWAS signals between NDVI of individual time points and the curve parameters. This probably can be explained by the generally no or low correlation between NDVI of individual time points with curve parameters.

Potential application of UAV-based NDVI measurements

NDVI is known to be associated with grain yield (Robert et al., 1999; Araus et al., 2001; Rojas, 2007; Spitkó et al., 2016). The correlation between NDVI and grain yield is weak in the early developmental stages, increases at tasseling stage, reaches maximum during mid-grain

filling period, and decreases after this period (Robert et al., 1999). And it has been shown that the correlation between NDVI and grain yield could be higher when estimated with UAV-based platforms (Kyratzis et al., 2017).

NDVI is correlated with grain yield and it can be measured remotely on a large number of selection candidates with UAV-based remote sensing platform. Therefore, UAV-based NDVI measurements can be used as secondary traits to indirectly select for grain yield and improve the genomic prediction model accuracy for grain yield. Indeed, a study in wheat has shown that NDVI measurements from aerial-based remote sensing platform used as secondary traits in genomic prediction models could increase prediction accuracy for grain yield (Rutkoski et al., 2016). NDVI growth curves incorporate developmental information across the whole growing season. It is possible that instead of the NDVI at a single time point, the whole NDVI growth curve is better correlated with grain yield. This can be tested by checking the correlation between grain yield and NDVI growth curve parameters in the future. If NDVI growth curve parameters do have high correlation with grain yield, these curve parameters can also be used as secondary traits to improve the prediction model accuracy for grain yield. The identified genetic loci of NDVI from GWAS in our research might also be incorporated into genomic selection models as diagnostic markers to select high yield varieties in the future.

Conclusions

This study conducted extensive analysis on NDVI data obtained from a UAV-HTPP in a large maize population. We conducted clustering analysis on time series NDVI data to understand the dynamics and the developmental trends of NDVI. We identified genotypic differences and dynamics changes of NDVI during different growth stages of maize. Our study also demonstrated the usefulness of time series NDVI data obtained from UAV-HTPPs in the GWAS analysis. We modeled the time series NDVI data with P-splines model and used the

model estimated curve parameters as phenotypic traits for the GWAS analysis. Our results showed the ability of UAV-based NDVI remote sensing for the genetic dissection of NDVI and the advantage of NDVI growth curve parameters over NDVI from individual time points for the detection of NDVI genetic loci.

Author Contributions

JY, KP, XL, and JW designed the study. JW, KP, XL, MD, and XY conducted the analyses. JW, XL, and JY wrote the manuscript with inputs from all authors.

Figures and Tables

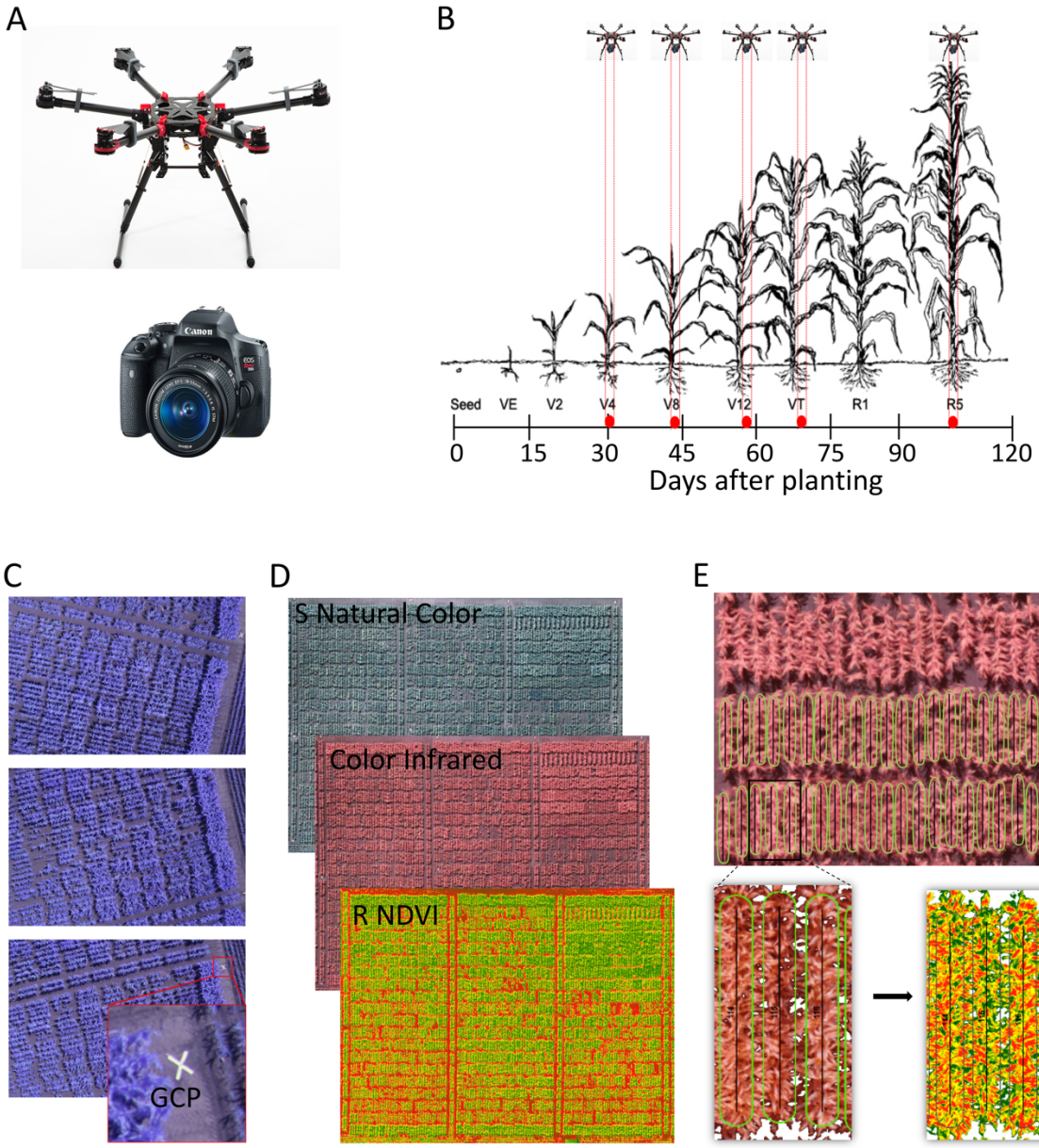


Figure 1. The UAV-HTPP system, the conducted UAV overflights around 5 growth stages and the image processing steps. (**A**) The UAV with a DJI S900 multirotor UAV and a Canon EOS Rebel SL1 color Infrared (CIR) converted camera. (**B**) Conducted 5 UAV overflights around 5 developmental stages of maize. (**C**) Three consecutive raw images from UAS overflight. GCP is shown by the white X in the bottom image. (**D**) S natural color, color infrared and R NDVI orthomosaic images generated from hundreds of raw images. (**E**) Example of the plot transect and 30 cm transect-buffer zone for plot level data extraction.

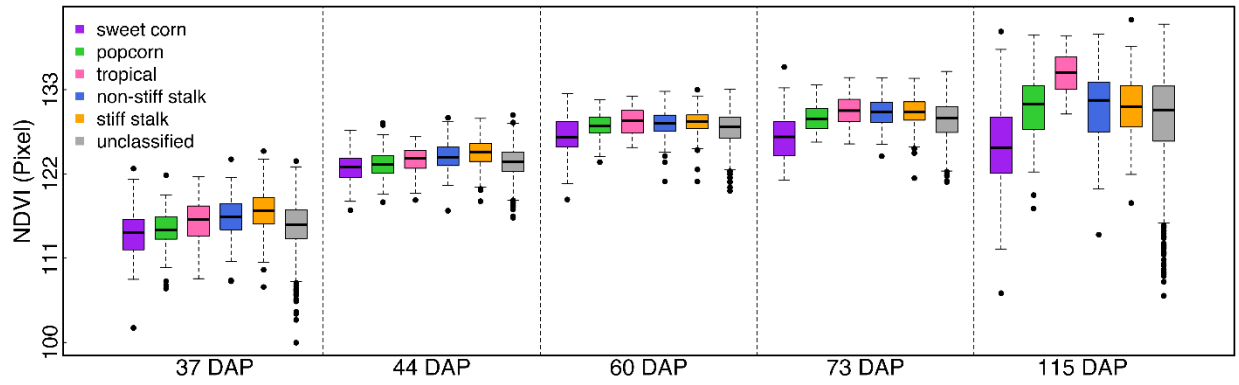


Figure 2. Distribution of NDVI at each of the 5 UAV overflights. NDVI distribution for each of the population group was plotted at each UAV overflights.

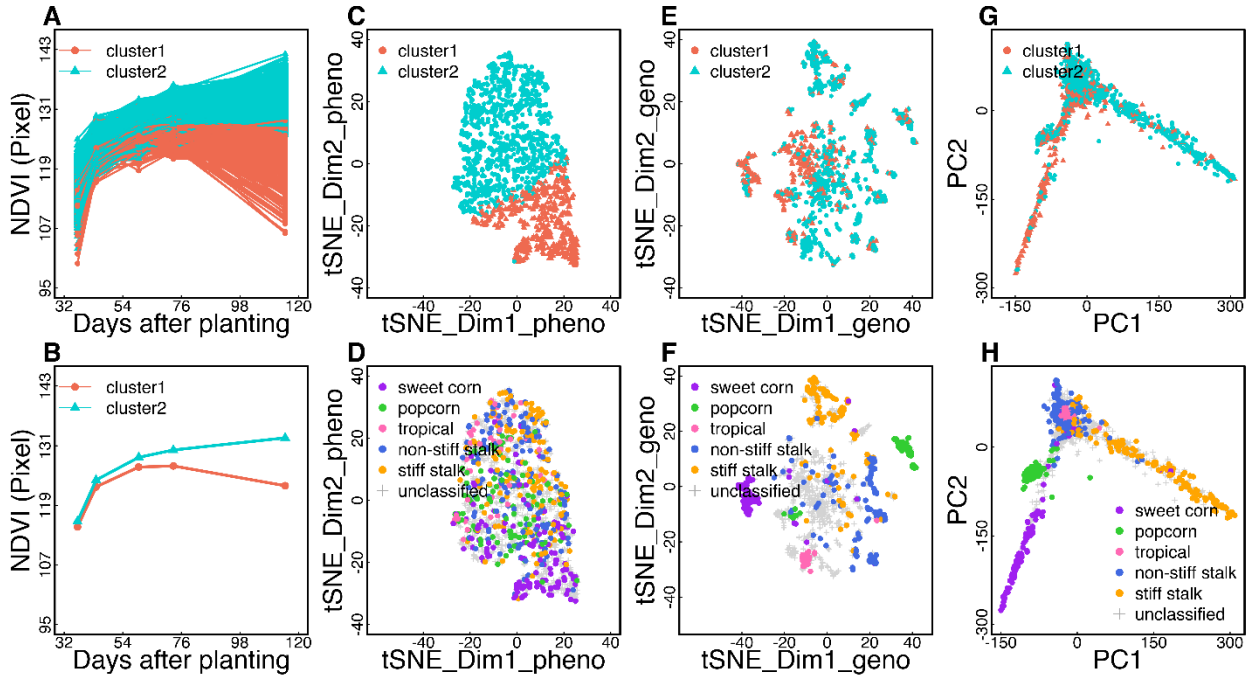


Figure 3. Clustering of time series NDVI and population structure analysis in 1752 maize accessions. (**A**) Growth curve of 1752 maize accessions with time series NDVI from 5 UAV overflights. Growth curve for each accession was colored according to the cluster which it belongs to. (**B**) Growth curve of each cluster with cluster average NDVI at each of the 5 overflights. (**C**) T-SNE with NDVI from 5 UAV overflights. Dot for each accession was colored according to the cluster which it belongs to. (**D**) T-SNE with NDVI from 5 UAV overflights. Dot for each accession was colored according to the population group which it belongs to. (**E**) T-SNE with 316,047 SNPs of 1752 accessions. Dot for each accession was colored according to the cluster which it belongs to. (**F**) T-SNE with 316,047 SNPs of 1752 accessions. Dot for each accession was colored according to the population group which it belongs to. (**G**) Plot of PC1 versus PC2. Dot for each accession was colored according to the which it belongs to. (H) Plot of PC1 versus PC2. Dot for each accession was colored according to population which it belongs to.

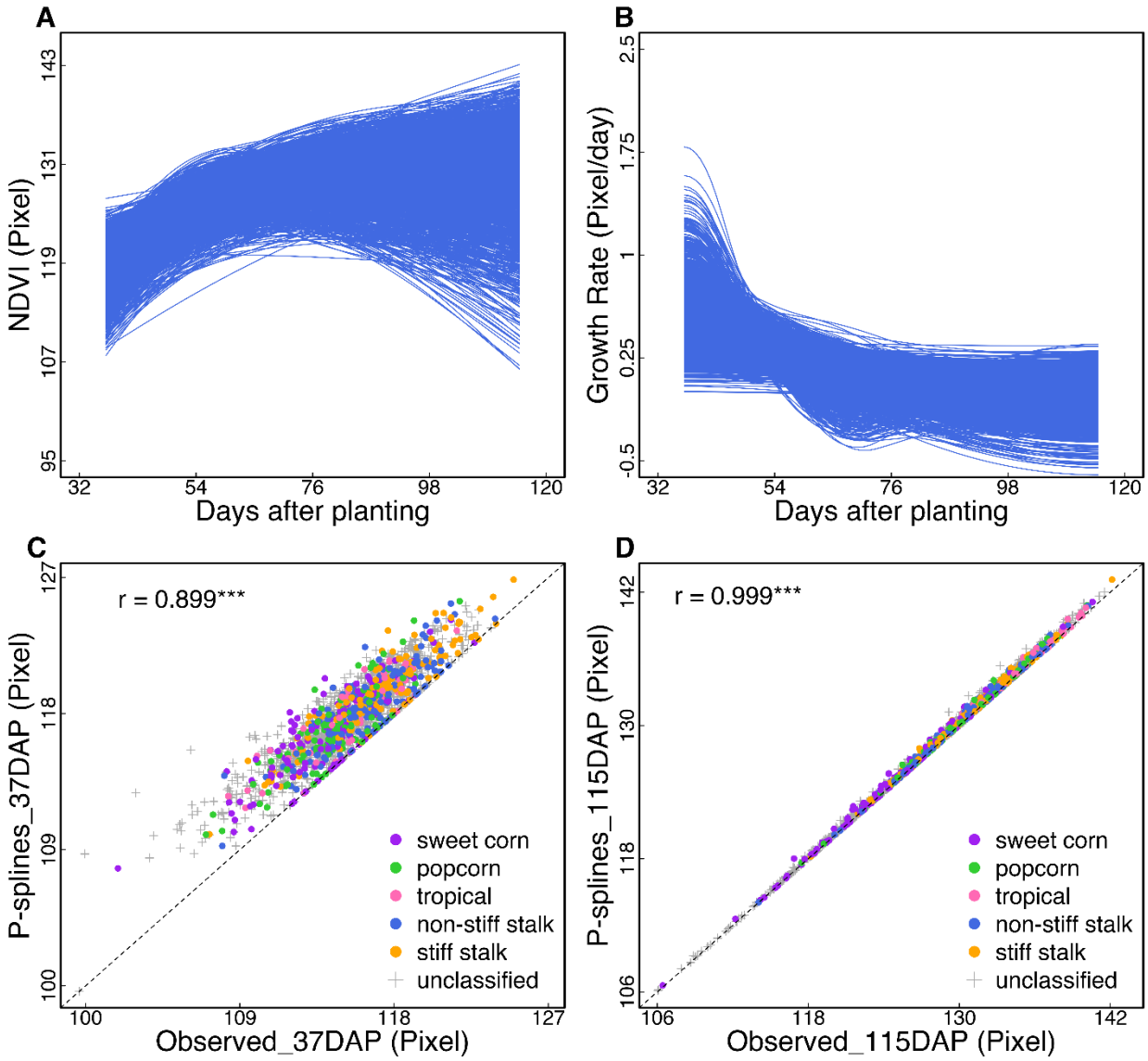


Figure 4. Modeling time series NDVI with P-splines. (**A**) Growth curve with P-splines model fitted NDVI. (**B**) P-splines model fitted growth rate across the growing season. (**C**) The correlation between P-splines model fitted NDVI with observed NDVI on 37 DAP. (**D**) The correlation between P-splines model fitted NDVI with observed NDVI on 115 DAP. ***$P <$ 0.001, **0.001 <$P <$ 0.01, *0.01 <$P <$ 0.05.

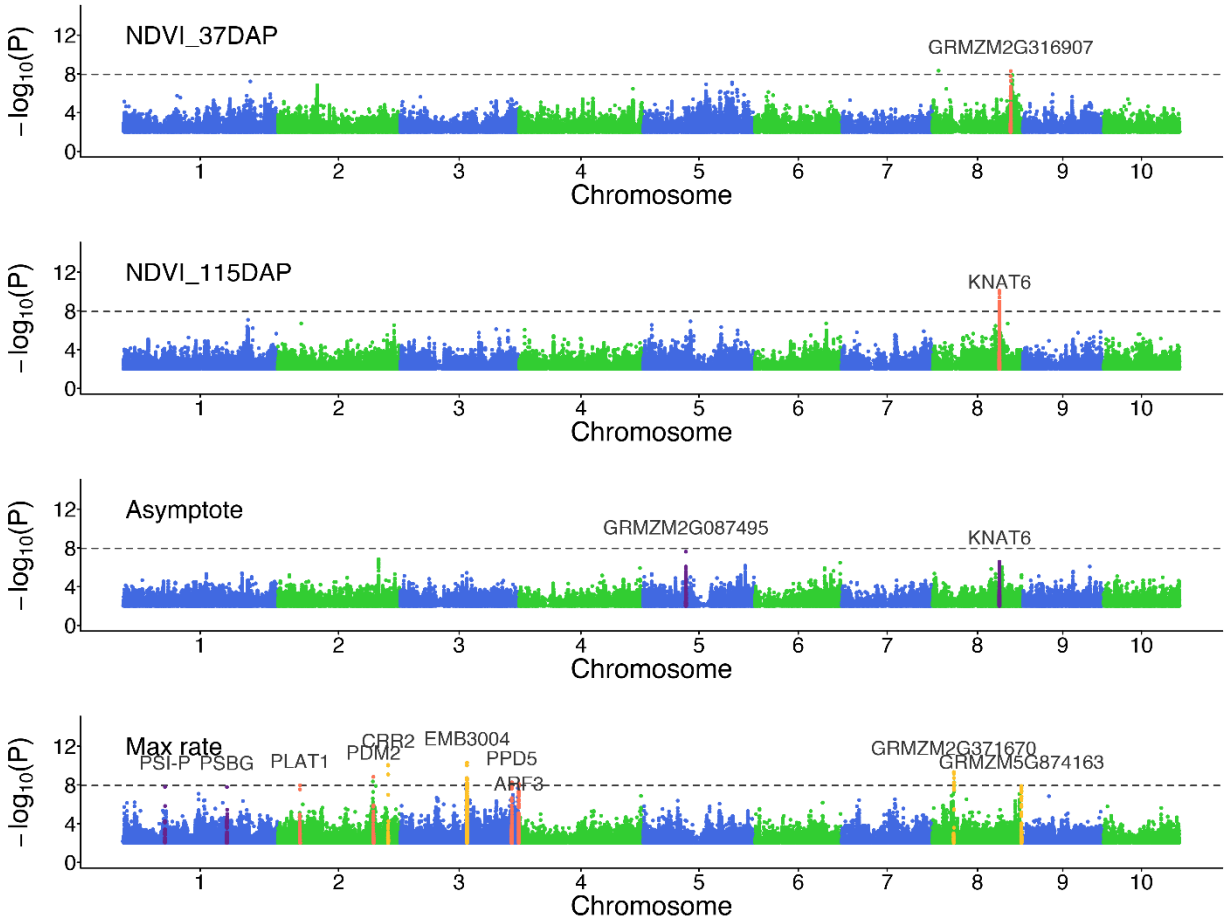


Figure 5. Genome-wide association mapping of NDVI and P-splines curve parameters. The horizontal line in each section is the Bonferroni-corrected significance threshold with 4,407,833 effective independent tests obtained from simpleM. The positions of plausible candidate genes and surrounding SNPs are indicated. When the tagged SNP of the gene is significantly associated with the trait and the gene is within 100 kb window surrounding the significantly associated SNPs, the surrounding SNPs of the gene is highlighted in coral color; when the tagged SNP of the gene is significantly associated with the trait and the gene is outside of the 100 kb window but within 1 Mb window surrounding the significantly associated SNPs, the surrounding SNPs of the gene is highlighted in orange color; when the tagged SNP of the gene is not significantly associated with the trait and the gene is within 100 kb window surrounding the association signal, the surrounding SNPs will be highlighted in purple color.

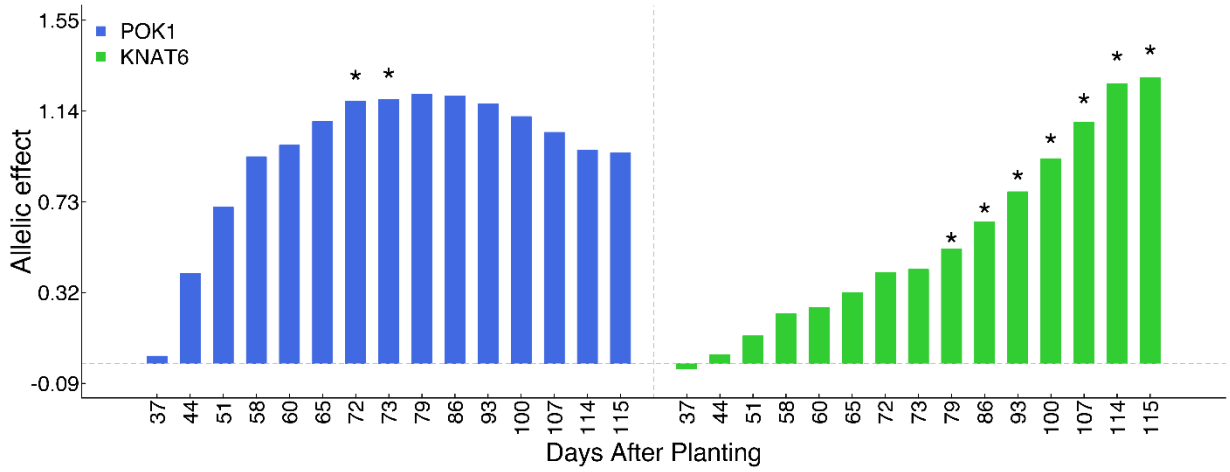


Figure 6. Dynamic changes of additive allelic effects of tagged SNP for potential candidate genes across growing season. Allelic effects for the most significantly associate SNP of the candidate gene *POK1* and *KNAT6* were obtained from the GWAS of P-splines fitted NDVI at 15 time points across the growing season. * indicates the tagged SNP for the gene is significantly associated with the trait at that specific time.

Table 1. Summary statistics for NDVI and manually measured flowering time, plant height, and ear height.

Trait	DAP	Pop group	Range	Mean	St.dev.	CV
NDVI	37	stiff stalk	107.23-124.98	117.18	3.09	2.63
		non-stiff stalk	107.98-123.9	116.41	2.60	2.24
		popcorn	107.03-121.83	114.65	2.50	2.18
		sweet corn	101.89-122.7	114.14	2.99	2.62
		tropical	108.33-121.67	115.53	3.10	2.68
		unclassified	99.64-123.67	115.24	3.13	2.72
		all 6 groups	**99.64-124.98**	**115.39**	**3.12**	**2.70**
	44	stiff stalk	118.43-129.24	124.80	1.91	1.53
		non-stiff stalk	117.19-129.33	124.19	1.91	1.53
		popcorn	118.31-128.64	123.29	1.83	1.49
		sweet corn	117.27-127.71	122.82	1.98	1.61
		tropical	118.59-126.85	123.73	2.01	1.62
		unclassified	106.5-129.68	123.55	2.06	1.67
		all 6 groups	**106.5-129.68**	**123.64**	**2.06**	**1.66**
	60	stiff stalk	121-132.98	128.68	1.69	1.31
		non-stiff stalk	121.01-132.77	128.50	1.80	1.40
		popcorn	123.53-131.68	128.24	1.57	1.22
		sweet corn	118.65-132.46	127.02	2.35	1.85
		tropical	125.45-132.13	128.90	1.78	1.38
		unclassified	119.77-133.05	127.93	2.02	1.58
		all 6 groups	**118.65-133.05**	**128.02**	**2.01**	**1.57**
	73	stiff stalk	121.45-134.48	130.09	1.97	1.52
		non-stiff stalk	124.31-134.58	130.05	1.83	1.41
		popcorn	126.15-133.61	129.28	1.62	1.26
		sweet corn	121.18-135.96	126.65	2.85	2.25
		tropical	125.92-134.59	130.39	2.07	1.59
		unclassified	120.94-135.39	129.04	2.51	1.94
		all 6 groups	**120.94-135.96**	**129.08**	**2.53**	**1.96**
	115	stiff stalk	118.2-142.14	130.74	3.92	3.00
		non-stiff stalk	114.06-140.22	130.81	4.30	3.28
		popcorn	117.47-140.17	130.57	3.98	3.05
		sweet corn	106.43-140.59	125.41	5.50	4.38
		tropical	129.81-140.06	135.15	2.84	2.10
		unclassified	106.06-141.55	129.20	5.99	4.63
		all 6 groups	**106.06-142.14**	**129.44**	**5.72**	**4.42**

Table 1 Continued

Trait	DAP	Pop group	Range	Mean	St.dev.	CV
Flowering time		stiff stalk	69-87	76.27	4.04	5.30
		non-stiff stalk	66-90	77.13	5.01	6.49
		popcorn	69-93	78.11	4.81	6.16
		sweet corn	56-98	70.90	6.34	8.94
		tropical	80-113	95.45	6.91	7.24
		unclassified	56-118	78.88	8.90	11.29
		all 6 groups	**56-118**	**78.29**	**8.64**	**11.03**
Plant height		stiff stalk	88-185	145.58	18.35	12.61
		non-stiff stalk	91-193	140.44	21.27	15.15
		popcorn	84-181	133.61	20.86	15.62
		sweet corn	49-169	106.69	25.53	23.93
		tropical	116-209	162.94	23.05	14.15
		unclassified	48-221	139.00	25.99	18.70
		all 6 groups	**48-221**	**137.50**	**26.54**	**19.30**
Ear height		stiff stalk	36-113	67.03	14.70	21.93
		non-stiff stalk	31-110	61.04	14.87	24.37
		popcorn	21-119	71.13	18.38	25.84
		sweet corn	5-94	38.71	17.51	45.25
		tropical	44-139	90.02	21.43	23.80
		unclassified	10-160	64.77	19.22	29.67
		all 6 groups	**5-160**	**63.71**	**20.26**	**31.79**

Table 2. K-means cluster result with NDVI from 5 UAV overflights.

Cluster	Pop group	Number of accessions	Percentage including unclassified	Percentage excluding unclassified
cluster 1	stiff stalk	29	5.07%	15.67%
	non-stiff stalk	36	6.29%	19.46%
	popcorn	24	4.20%	12.97%
	sweet corn	96	16.78%	51.89%
	tropical	0	0.00%	0.00%
	unclassified	387	67.65%	
	all 6 groups	572	100.00%	
cluster 2	stiff stalk	120	10.17%	29.70%
	non-stiff stalk	115	9.75%	28.47%
	popcorn	84	7.12%	20.79%
	sweet corn	38	3.22%	9.41%
	tropical	47	3.98%	11.63%
	unclassified	776	65.76%	
	all 6 groups	1180	100.00%	

Supplementary Information

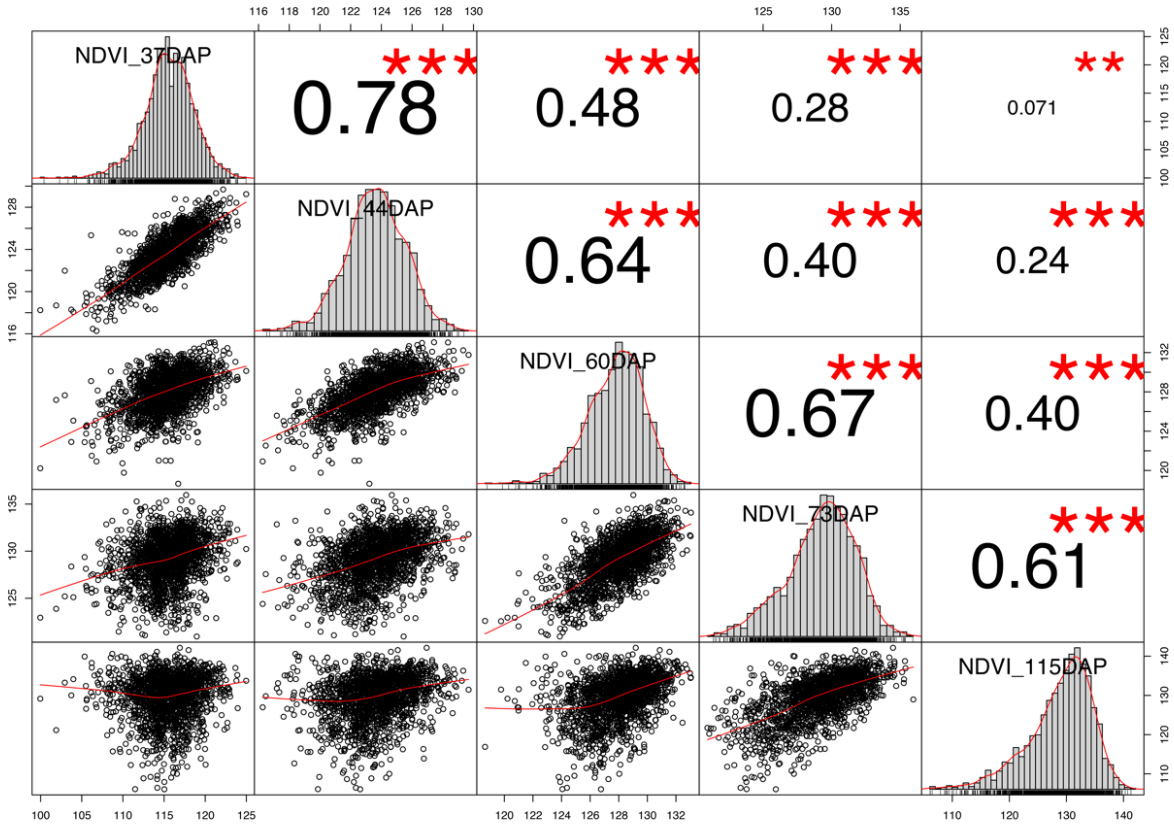


Figure S1. Distributions and correlations for NDVIs from 5 UAV overflights. ***$P < 0.001$, **$0.001 < P < 0.01$, *$0.01 < P < 0.05$.

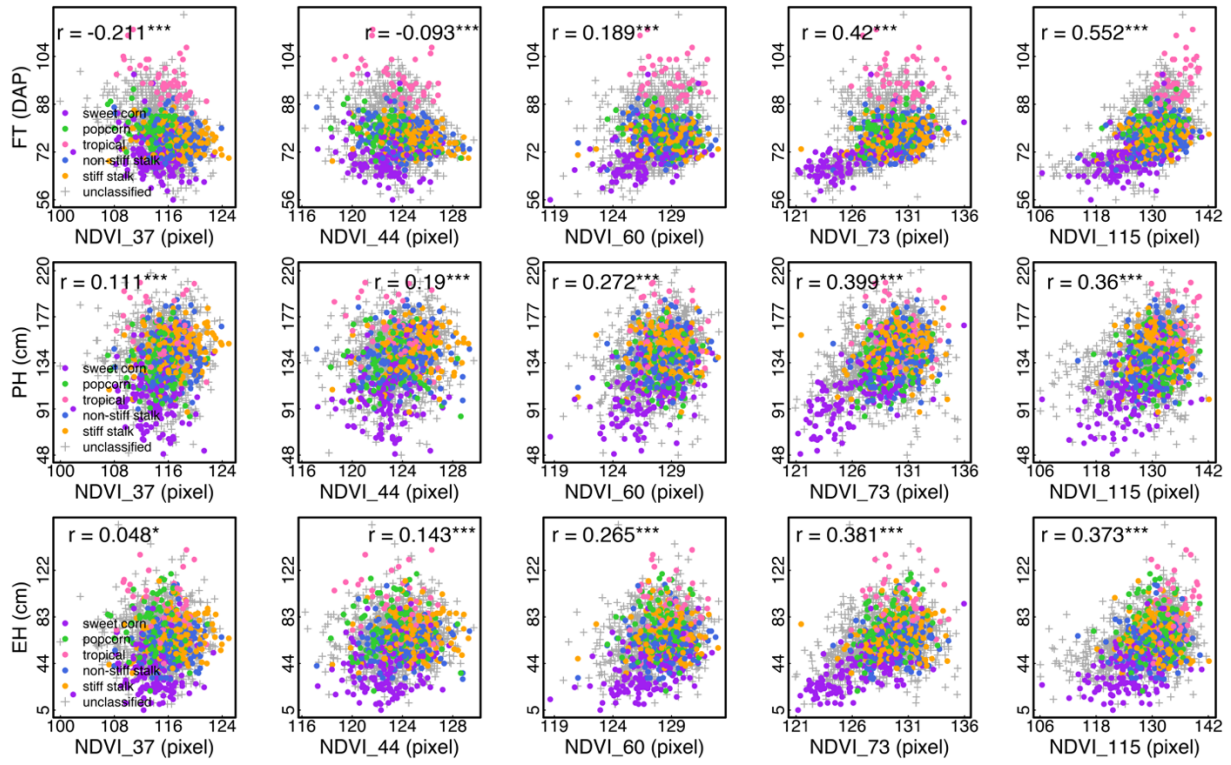


Figure S2. Pearson correlation between NDVI from 5 UAV overflights and manually measured flowering time, plant height, and ear height. ***$P < 0.001$, **$0.001 < P < 0.01$, *$0.01 < P < 0.05$. FT for flowering time, PH for plant height, and EH for ear height.

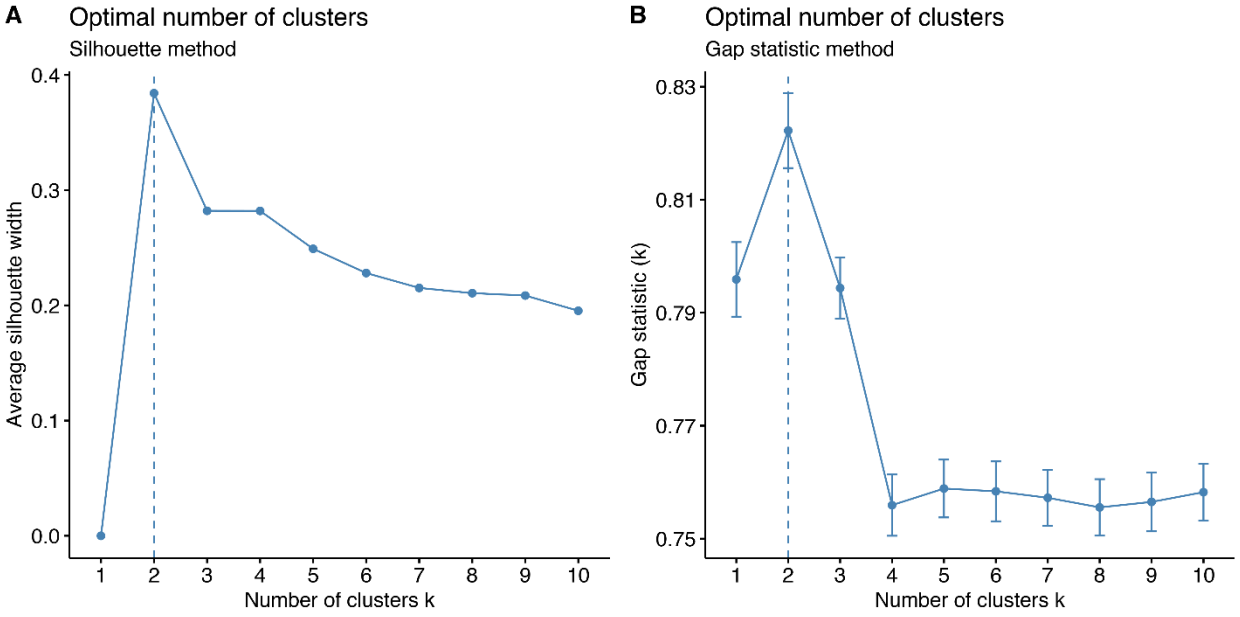


Figure S3. Optimum number of clusters with silhouette method (A) and gap statistics method (B).

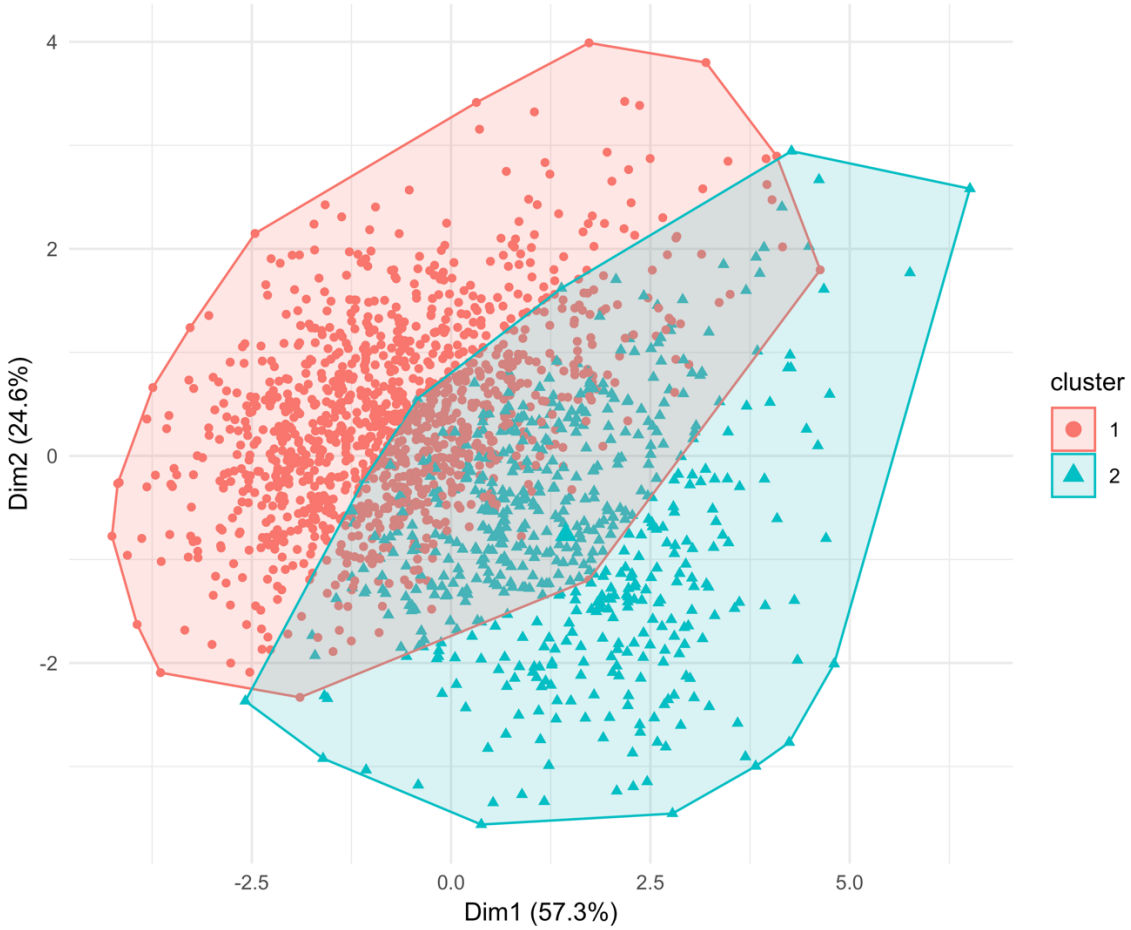


Figure S4. K-means clustering of time series NDVI from 5 UAV overflights.

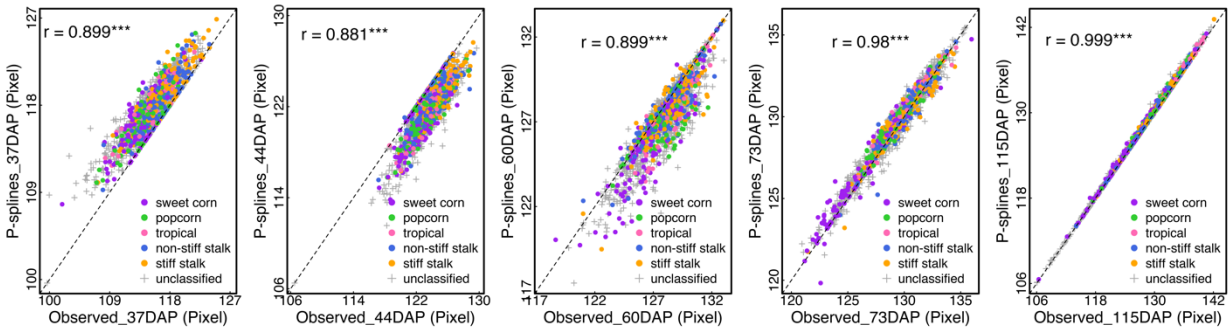


Figure S5. Correlation between observed NDVI and P-splines fitted NDVI. ***$P < 0.001$, **$0.001 < P < 0.01$, *$0.01 < P < 0.05$.

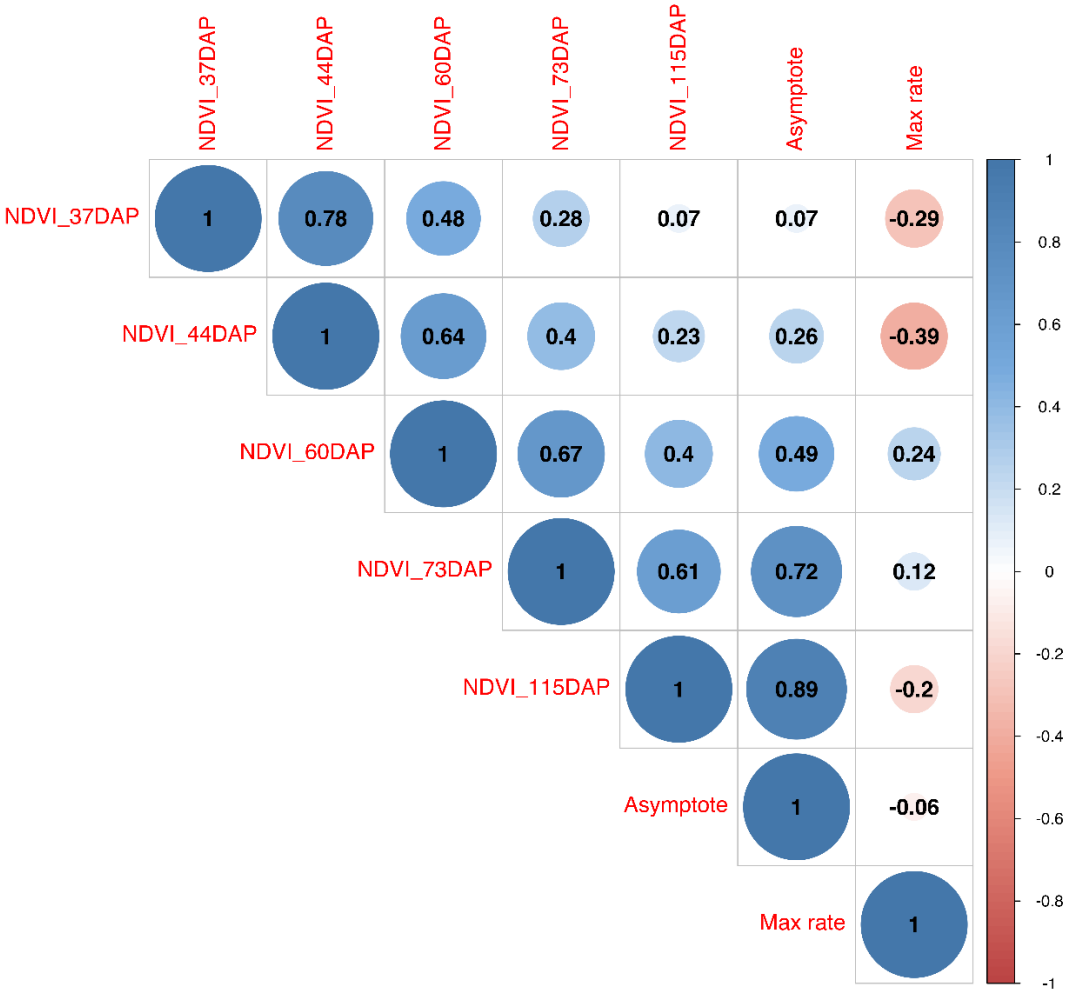


Figure S6. Pairwise Pearson correlation between NDVI from individual growth stages and P-splines curve parameters. The size and shade of each circle represent the strength of each relationship. It means the relationship is not significant if there are no circle for it. Blue color indicates the correlation is positive, while red color indicates the correlation is negative.

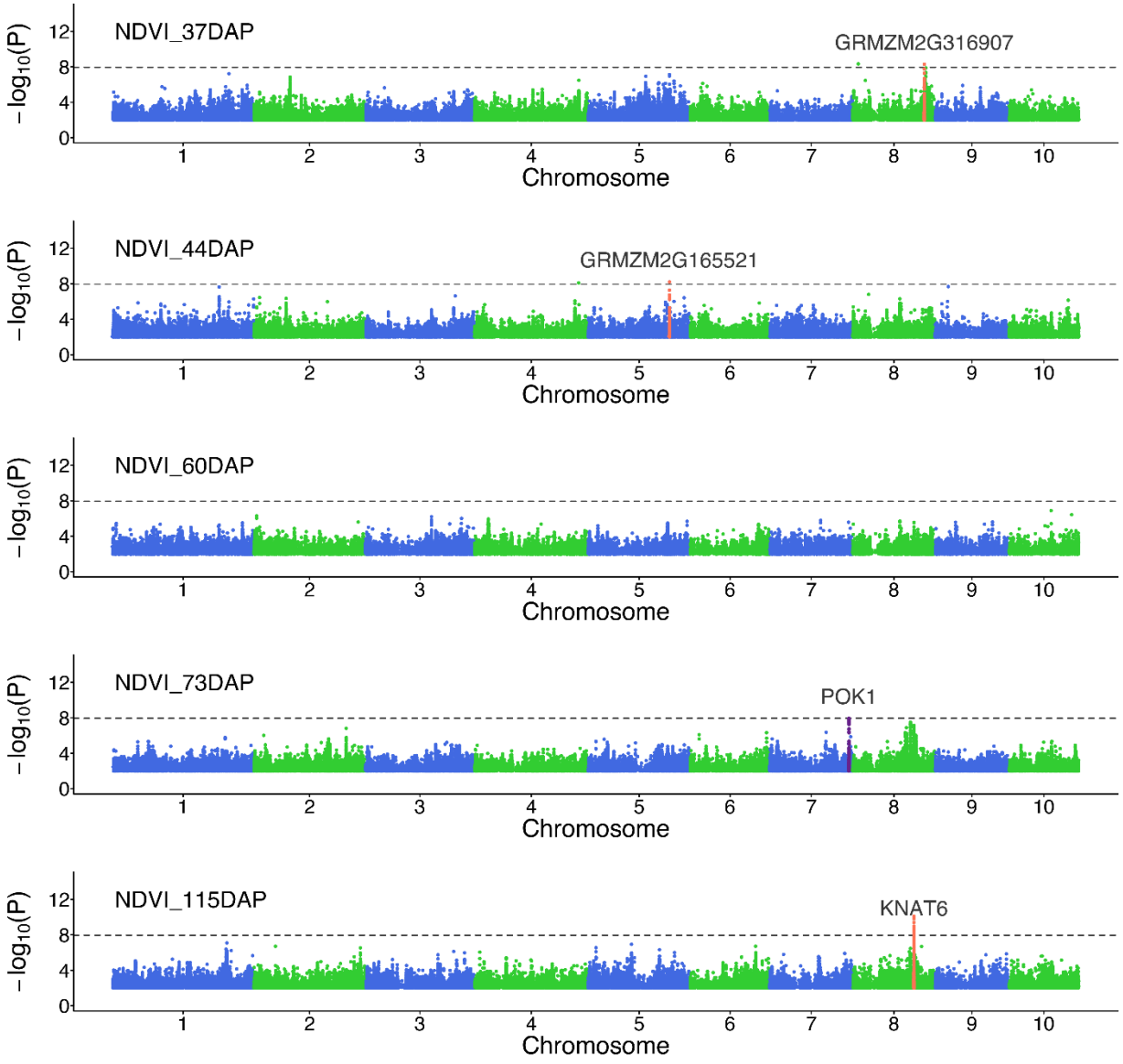


Figure S7. Genome-wide association mapping of NDVI from 5 UAV overflights. The horizontal line in each section is the Bonferroni-corrected significance threshold with 4,407,833 effective independent tests obtained from simpleM. The positions of plausible candidate genes and surrounding SNPs are indicated. When the tagged SNP of the gene is significantly associated with the trait and the gene is within 100 kb window surrounding the significantly associated SNPs, the surrounding SNPs of the gene is highlighted in coral color; when the tagged SNP of the gene is not significantly associated with the trait and the gene is within 100 kb window surrounding the association signal, the surrounding SNPs is highlighted in purple color.

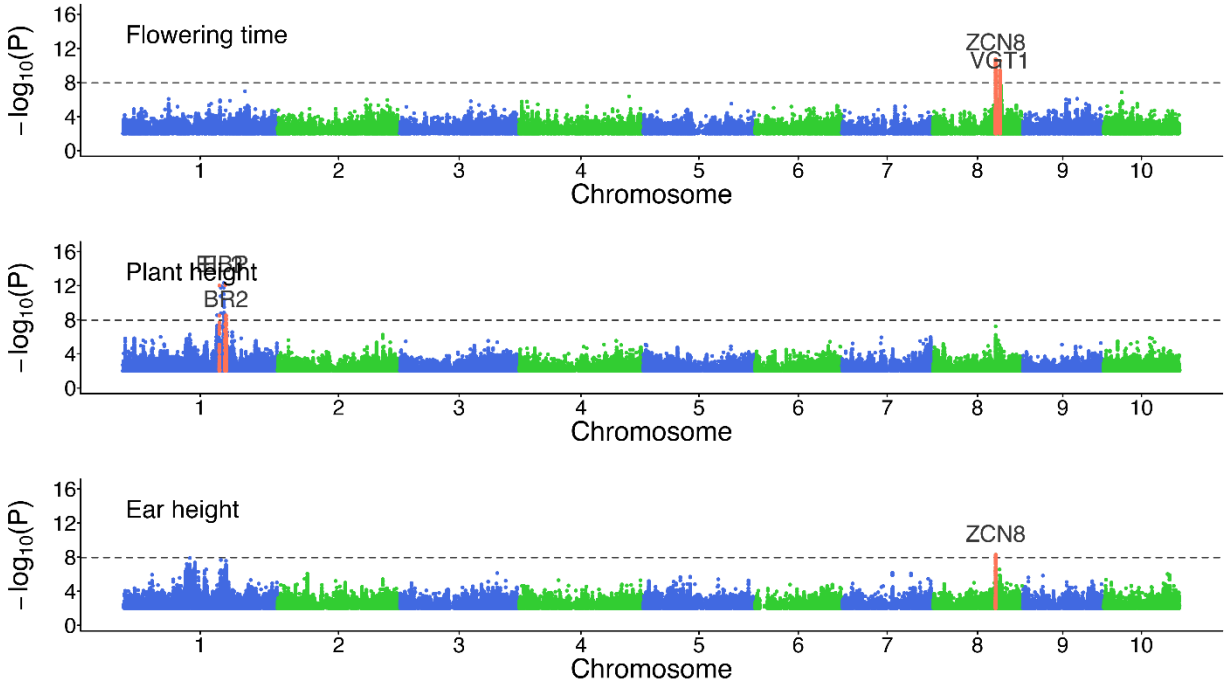
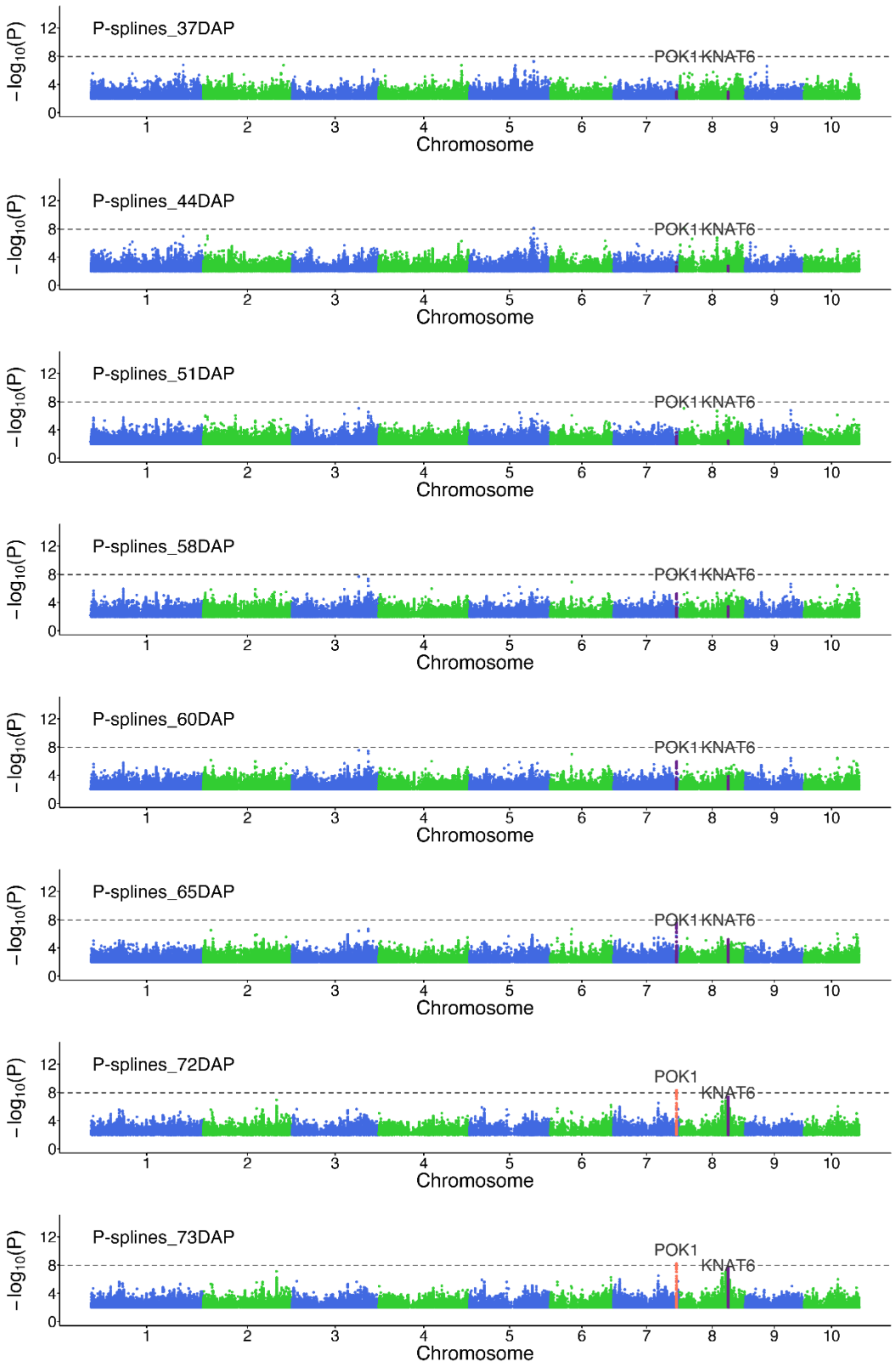


Figure S8. Genome-wide association mapping of flowering time, plant height, and ear height. The horizontal line in each section is the Bonferroni-corrected significance threshold with 4,407,833 effective independent tests obtained from simpleM. The positions of plausible candidate genes and surrounding SNPs are indicated.


P-splines_37DAP
P-splines_44DAP
P-splines_51DAP
P-splines_58DAP
P-splines_60DAP
P-splines_65DAP
P-splines_72DAP
P-splines_73DAP
POK1
KNAT6
−log10(P)
Chromosome

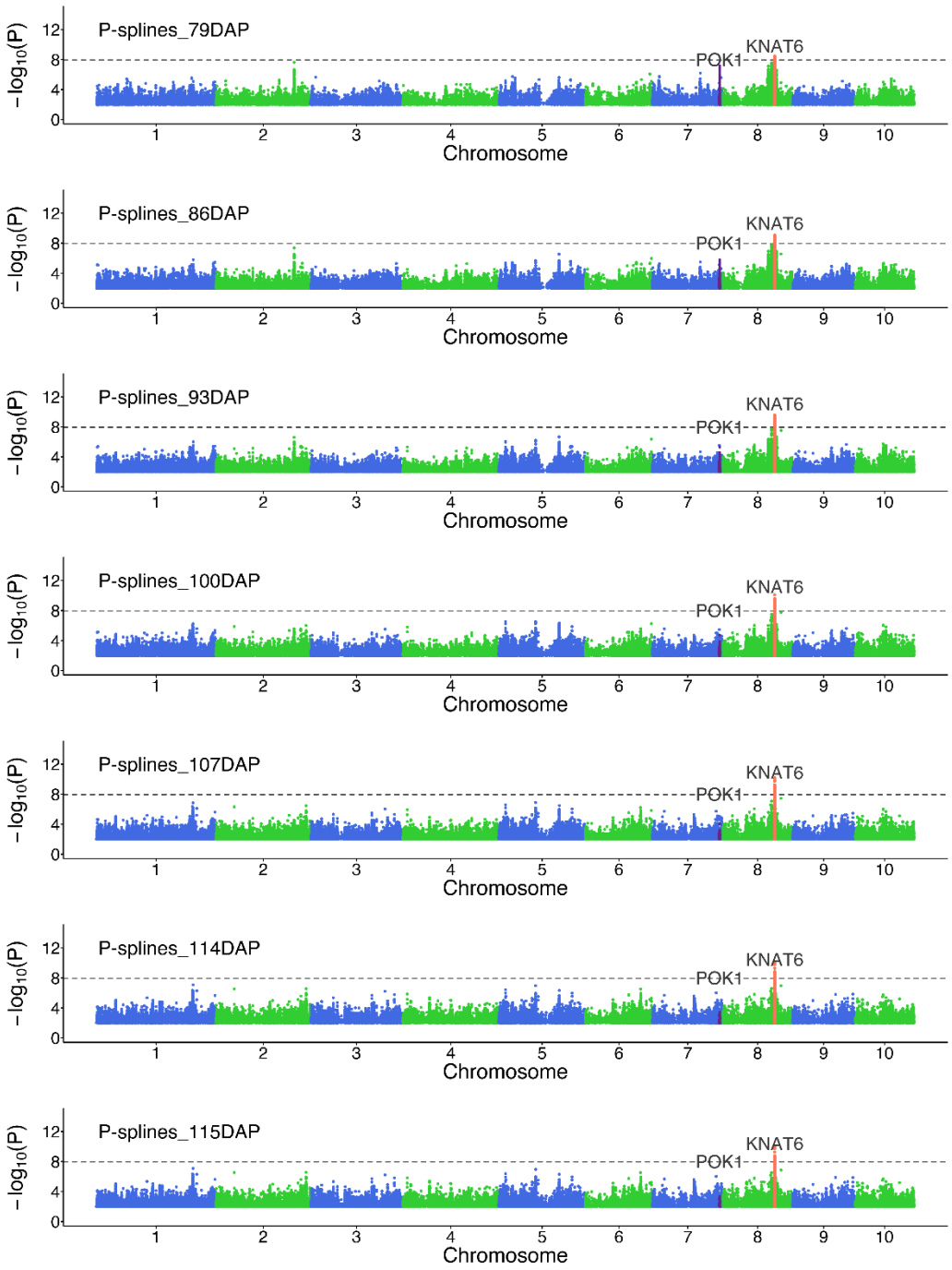


Figure S9. Genome-wide association mapping of 15 P-splines model fitted NDVI. The horizontal line in each section is the Bonferroni-corrected significance threshold with 4,407,833 effective independent tests obtained from simpleM. The positions of plausible candidate genes and surrounding SNPs are indicated. When the tagged SNP of the gene is significantly associated with the trait and the gene is within 100 kb window surrounding the significantly associated SNPs, the surrounding SNPs of the gene is highlighted in coral color; when the tagged SNP of the gene is not significantly associated with the trait and the gene is within 100 kb window surrounding the association signal, the surrounding SNPs is highlighted in purple color.

Table S1. Summary statistics for NDVI of each cluster.

Cluster	DAP	Range	Mean	SD	CV
cluster1	37	99.96-122.7	114.66	2.89	2.52
	44	116.27-128.64	122.72	1.82	1.48
	60	118.65-132.46	126.70	1.88	1.48
	73	120.94-132.53	126.93	2.29	1.80
	115	106.06-131.48	122.95	4.50	3.66
cluster2	37	99.64-124.98	115.75	3.16	2.73
	44	106.5-129.68	124.09	2.02	1.63
	60	121.01-133.05	128.65	1.75	1.36
	73	123.2-135.96	130.13	1.89	1.46
	115	125.83-142.14	132.58	2.90	2.19

Table S2. GWAS detected genes within 100 kb from the associated SNPs.

Trait	Win (Kb)	Gene	Chr	Start	End	Function	Alias
NDVI_44DAP	1	GRMZM2G165521	5	173,455,046	173,459,109	TPR-like superfamily protein	_
NDVI_115DAP	1	GRMZM2G094241	8	130,383,323	130,388,890	KNOTTED1-like homeobox gene 6	*KNAT6*
NDVI_115DAP	50	GRMZM2G034043	8	130,259,650	130,263,976	TPR-like superfamily protein	*TPR5*
NDVI_115DAP	100	GRMZM2G020096	8	130,459,536	130,471,342	glutamate-cysteine ligase, chloroplast precursor	*GSH1*
max rate	1	GRMZM2G114399	3	218,573,327	218,576,745	photosystem II reaction center PsbP family protein	*PPD5*
max rate	10	GRMZM2G002043	2	186,771,213	186,777,030	Pentatricopeptide repeat (PPR) superfamily protein	*PDM2*
max rate	50	GRMZM2G361376	3	232,108,213	232,112,166	auxin-responsive factor AUX/IAA-related	*ARF3*
max rate	100	GRMZM2G018275	2	43,746,150	43,747,663	Lipase/lipooxygenase, PLAT/LH2 family protein	*PLAT1*
P-splines_72DAP	5	GRMZM2G300709	7	168,720,518	168,741,026	phragmoplast orienting kinesin 1	*POK1*
P-splines _73DAP	5	GRMZM2G300709	7	168,720,518	168,741,026	phragmoplast orienting kinesin 1	*POK1*
P-splines 79DAP	5	GRMZM2G094241	8	130,383,323	130,388,890	KNOTTED1-like homeobox gene 6	*KNAT6*
P-splines _86DAP	5	GRMZM2G094241	8	130,383,323	130,388,890	KNOTTED1-like homeobox gene 6	*KNAT6*
P-splines _93DAP	5	GRMZM2G094241	8	130,383,323	130,388,890	KNOTTED1-like homeobox gene 6	*KNAT6*
P-splines _100DAP	5	GRMZM2G094241	8	130,383,323	130,388,890	KNOTTED1-like homeobox gene 6	*KNAT6*
P-splines _107DAP	5	GRMZM2G094241	8	130,383,323	130,388,890	KNOTTED1-like homeobox gene 6	*KNAT6*
P-splines _114DAP	5	GRMZM2G094241	8	130,383,323	130,388,890	KNOTTED1-like homeobox gene 6	*KNAT6*
P-splines _115DAP	5	GRMZM2G094241	8	130,383,323	130,388,890	KNOTTED1-like homeobox gene 6	*KNAT6*
Flowering time	1	GRMZM2G700665	8	131,576,889	131,580,316	target of early activation tagged (EAT) 2	*VGT1*
Flowering time	100	GRMZM2G179264	8	123,030,387	123,032,135	PEBP (phosphatidylethanolamine-binding protein) family protein	*ZCN8*
Plant height	5	GRMZM2G369472	1	199,674,463	199,675,600	ethylene-responsive element binding protein	*EBP*
Ear height	10	GRMZM2G179264	8	123,030,387	123,032,135	PEBP (phosphatidylethanolamine-binding protein) family protein	*ZCN8*

References

Andrade-Sanchez P, Gore MA, Heun JT, Thorp KR, Carmo-Silva AE, French AN, Salvucci ME, White JW (2014) Development and evaluation of a field-based high-throughput phenotyping platform. Funct Plant Biol 41: 68–79

Araus JL, Casadesus J, Bort1 J (2001) Recent tools for the screening of physiological traits determining yield. CIMMYT, Mexico

Ballesteros R, Ortega JF, Hernández D, Moreno MA (2014) Applications of georeferenced high-resolution images obtained with unmanned aerial vehicles. Part I: Description of image acquisition and processing. Precis Agric 15: 579–592

Berger B, Parent B, Tester M (2010) High-throughput shoot imaging to study drought responses. J Exp Bot 61: 3519–3528

Berni JAJ, Zarco-Tejada PJ, Suárez L, Fereres E (2009) Thermal and narrowband multispectral remote sensing for vegetation monitoring from an unmanned aerial vehicle. IEEE Trans Geosci Remote Sens 47: 722–738

Bort J, Casadesus J, Nachit MM, Araus JL (2005) Factors affecting the grain yield predicting attributes of spectral reflectance indices in durum wheat: growing conditions, genotype variability and date of measurement. Int J Remote Sens 26: 2337-2358

Calderon CP, Martinez JG, Carroll RJ, Sorensen DC (2010) P-Splines using derivative information. Multiscale Model Simul 8: 1562-1580

Chao HY, Cao YC, Chen YQ (2010) Autopilots for small unmanned aerial vehicles: A survey. Int J Control Autom Syst 8: 36–44

Chapman SC, Merz T, Chan A, Jackway P, Hrabar S, Dreccer MF, Holland E, Zheng B, Ling TJ, Jimenez-Berni J (2014) Pheno-copter: a low-altitude, autonomous remote-sensing robotic helicopter for high-throughput field-based phenotyping. Agronomy 4: 279–301

Christopher JT, Christopher MJ, Borrell AK, Fletcher S, Chenu K (2016) Stay-green traits to improve wheat adaptation in well-watered and water-limited environments. J Exp Bot 67: 5159-5172

Cobb JN, DeClerck G, Greenberg A, Clark R, McCouch S (2013) Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype-phenotype relationships and its relevance to crop improvement. Theor Appl Genet 126: 867–887

Condorelli GE, Maccaferri M, Newcomb M, Andrade-Sanchez P, White JW, French AN, Sciara G, Ward R, Tuberosa R (2018) Comparative aerial and ground based high throughput phenotyping for the genetic dissection of NDVI as a proxy for drought adaptive traits in durum wheat. Front Plant Sci 9: 893

Du L, Zhang J, Qu S, Zhao Y, Su B, Lv X, Li R, Wan Y, Xiao J (2017) The pentratricopeptide repeat protein pigment-defective mutant2 is involved in the regulation of chloroplast development and chloroplast gene expression in Arabidopsis. Plant Cell Physiol 58: 747-759

Duncan WG, Williams WA, Loomis RS (1967) Tassels and the productivity of maize 1. Crop Sci 7: 37-39

Dunford R, Michel K, Gagnage M, Piegay H, Tremelo ML (2009) Potential and constraints of unmanned aerial vehicle technology for the characterization of Mediterranean riparian forest. Int J Remote Sens 30: 4915–4935

Eilers PHC, Marx BD (1996) Flexible smoothing with B-splines and penalties. Stat Sci 11: 89-102

Fernandez MGS, Bao Y, Tang L, Schnable PS (2017) A high-throughput, field-based phenotyping technology for tall biomass crops. Plant Physiol 174: 2008–2022

Fiorani F, Schurr U (2013) Future scenarios for plant phenotyping. Annu Rev Plant Biol 64: 267–291

Furbank RT, Tester M (2011) Phenomics - technologies to relieve the phenotyping bottleneck. Trends Plant Sci 16: 635–644

Gao X (2011) Multiple testing corrections for imputed SNPs. Genet Epidemiol 35: 154-158

Gao X, Becker LC, Becker DM, Starmer JD, Province MA (2010) Avoiding the high Bonferroni penalty in genome-wide association studies. Genet Epidemiol 34: 100-105

Gao X, Starmer J, Martin ER (2008) A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. Genet Epidemiol 32: 361-369

Gerland P, Raftery AE, Sevcikova H, Li N, Gu D, Spoorenberg T, Alkema L, Fosdick BK, Chunn J, Lalic N, et al (2014) World population stabilization unlikely this century. Science 346: 234–237

Govaerts B, Verhulst N (2010) The normalized difference vegetation index (NDVI) Greenseeker(TM) handheld sensor: toward integrated evaluation of crop management part B-user guide. CIMMYT, Mexico

Haghighattalab A, Perez LG, Mondal S, Singh D, Schinstock D, Rutkoski J, Ortiz-Monasterio I, Singh RP, Goodin D, Poland J (2016) Application of unmanned aerial systems for high throughput phenotyping of large wheat breeding nurseries. Plant Methods 12: 35

Han L, Yang G, Yang H, Xu B, Li Z, Yang X (2018) Clustering field-based maize phenotyping of plant-height growth and canopy spectral dynamics using a UAV remote-sensing approach. Front Plant Sci 9: 1638

Hunt ER, Cavigelli M, Daughtry CST, Mcmurtrey JE, Walthall CL (2005) Evaluation of digital photography from model aircraft for remote sensing of crop biomass and nitrogen status. Precis Agric 6: 359–378

Hurtado PX, Schnabel SK, Zaban A, Veteläinen M, Virtanen E, Eilers PHC, van Eeuwijk FA, Visser RGF, Maliepaard C (2012) Dynamics of senescence-related QTLs in potato. Euphytica 183: 289-302

Kaufman L, Rousseeuw PJ (1990) Finding Groups in Data: An Introduction to Cluster Analysis (Wiley Series in Probability and Statistics). John Wiley & Sons, Inc, Hoboken, New Jersey, USA

Kumar R, Silva L (1973) Light ray tracing through a leaf cross section. Appl Opt 12: 2950-2954

Kyratzis AC, Skarlatos DP, Menexes GC, Vamvakousis VF, Katsiotis A (2017) Assessment of vegetation indices derived by UAV imagery for durum wheat phenotyping under a water limited and heat stressed Mediterranean environment. Front Plant Sci 8: 1114

Lewis JE, Rowland J, Nadeau A (1998) Estimating maize production in Kenya using NDVI: some statistical considerations. Int J Remote Sens 19: 2609-2617

Liebisch F, Kirchgessner N, Schneider D, Walter A, Hund A (2015) Remote, aerial phenotyping of maize traits with a mobile multi-sensor approach. Plant Methods 11: 9

Lincoln C, Long J, Yamaguchi J, Serikawa K, Hake S (1994) A knotted1-like homeobox gene in Arabidopsis is expressed in the vegetative meristem and dramatically alters leaf morphology when overexpressed in transgenic plants. Plant Cell 6: 1859-1876

Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, Gore MA, Buckler ES, Zhang Z (2012) GAPIT: genome association and prediction integrated tool. Bioinformatics 28: 2397–2399

Lipka E, Gadeyne A, Stöckle D, Zimmermann S, De Jaeger G, Ehrhardt DW, Kirik V, Van Damme D, Müller S (2014) The phragmoplast-orienting kinesin-12 class proteins translate the positional information of the preprophase band to establish the cortical division zone in Arabidopsis thaliana. Plant Cell 26: 2617-2632

Van Der Maaten L, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9: 2579-2605

Marti J, Bort J, Slafer GA, Araus JL (2007) Can wheat yield be assessed by early measurements of normalized difference vegetation index? Ann Appl Biol 50: 253-257

Multani DS, Briggs SP, Chamberlin MA, Blakeslee JJ, Murphy AS, Johal GS (2003) Loss of an MDR transporter in compact stalks of maize br2 and sorghum dw3 mutants. Science 302: 81-84

Ray DK, Mueller ND, West PC, Foley JA (2013) Yield trends are insufficient to double global crop production by 2050. PLoS One 8: e66428

Rojas O (2007) Operational maize yield model development and validation based on remote sensing and agro-meteorological data in Kenya. Int J Remote Sens 28: 3775-3793

Romay MC, Millard MJ, Glaubitz JC, Peiffer JA, Swarts KL, Casstevens TM, Elshire RJ, Acharya CB, Mitchell SE, Flint-Garcia SA (2013) Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biol 14: R55

Romero Navarro JA, Willcox M, Burgueño J, Romay C, Swarts K, Trachsel S, Preciado E, Terron A, Delgado HV, Vidal V, et al (2017) A study of allelic diversity underlying flowering-time adaptation in maize landraces. Nat Genet 49: 476-480

Roose JL, Frankel LK, Bricker TM (2011) Developmental defects in mutants of the PsbP domain protein 5 in Arabidopsis thaliana. PLoS One 6: e28624

Rundquist DC, Narumalani S, Narayanan RM (2001) A review of wetlands remote sensing and defining new considerations. Remote Sens Rev 20: 207-226

Rutkoski J, Poland J, Mondal S, Autrique E, Pérez LG, Crossa J, Reynolds M, Singh R (2016) Canopy temperature and vegetation indices from high-throughput phenotyping improve accuracy of pedigree and genomic selection for grain yield in wheat. G3 6: 2799-2808

Salvi S, Sponza G, Morgante M, Tomes D, Niu X, Fengler KA, Meeley R, Ananiev E V., Svitashev S, Bruggemann E, et al (2007) Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. Proc Natl Acad Sci U S A 104: 11376-11381

Schippers JHM, Schmidt R, Wagstaff C, Jing HC (2015) Living to die and dying to live: the survival strategy behind leaf senescence. Plant Physiol 169: 914-930

Spitkó T, Nagy Z, Zsubori ZT, Szőke C, Berzy T, Pintér J, Marton CL (2016) Connection between normalized difference vegetation index and yield in maize. Plant, Soil Environ 7: 293-298

Tibshirani R, Walther G, Hastie T (2001) Estimating the number of clusters in a data set via the gap statistic. J R Stat Soc Ser B Stat Methodol 63: 411-423

Torii KU (2004) Leucine-rich repeat receptor kinases in plants: structure, function, and signal transduction pathways. Int Rev Cytol 234: 1-46

Troyer AF, Larkins JR (1985) Selection for early flowering in corn: 10 late synthetics. Crop Sci 25: 695-697

Vadivambal R, Jayas DS (2011) Applications of thermal imaging in agriculture and food industry-a review. Food Bioprocess Technol 4: 186–199

Walter A, Liebisch F, Hund A (2015) Plant phenotyping: from bean weighing to image analysis. Plant Methods 11: 14

Wang R, Cherkauer K, Bowling L (2016) Corn response to climate stress detected with satellite-based NDVI time series. Remote Sens 8: 269

White JW, Andrade-Sanchez P, Gore MA, Bronson KF, Coffelt TA, Conley MM, Feldmann KA, French AN, Heun JT, Hunsaker DJ, et al (2012) Field-based phenomics for plant genetics research. F Crop Res 133: 101–112

Wu C, Niu Z, Tang Q, Huang W (2008) Estimating chlorophyll content from hyperspectral vegetation indices: Modeling and validation. Agric For Meteorol 148: 1230–1241

Xing A, Gao Y, Ye L, Zhang W, Cai L, Ching A, Llaca V, Johnson B, Liu L, Yang X, et al (2015) A rare SNP mutation in Brachytic2 moderately reduces plant height and increases yield potential in maize. J Exp Bot 66: 3791-3802

Yang GJ, Liu JG, Zhao CJ, Li ZH, Huang YB, Yu HY, Xu B, Yang XD, Zhu DM, Zhang XY, et al (2017) Unmanned aerial vehicle remote sensing for field-based crop phenotyping: current status and perspectives. Front Plant Sci 8: 1111

Yonah IB, Mourice SK, Tumbo SD, Mbilinyi BP, Dempewolf J (2018) Unmanned aerial vehicle-based remote sensing in monitoring smallholder, heterogeneous crop fields in Tanzania. Int J Remote Sens 39: 5453-5471

Yu J, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38: 203-208

Zaman-Allah M, Vergara O, Araus JL, Tarekegne A, Magorokosho C, Zarco-Tejada PJ, Hornero A, Albà AH, Das B, Craufurd P, et al (2015) Unmanned aerial platform-based multi-spectral imaging for field phenotyping of maize. Plant Methods 11: 35

Zhang C, Kovacs JM (2012) The application of small unmanned aerial systems for precision agriculture: a review. Precis Agric 13: 693–712

Zhang Z, Ersoz E, Lai C-Q, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordovas JM (2010) Mixed linear model approach adapted for genome-wide association studies. Nat Genet 42: 355-360

Zia S, Romano G, Spreer W, Sanchez C, Cairns J, Araus JL, Müller J (2013) Infrared thermal imaging as a rapid tool for identifying water‐stress tolerant maize genotypes of different phenology. J Agron Crop Sci 199: 75–84

CHAPTER 4. GENERAL CONCLUSION

Domesticated accessions have higher [AT] than wild accessions at the dynamic part of the genome, which suggests the important effect of AT-bias mutation on the overall pattern of base-composition variation. The difference in base composition between domesticated accessions and wild accessions varies between different parts of the genome, with non-genic part having larger [AT]-difference than genic SNPs and pericentromeric regions having larger [AT]-difference than chromosome arms. The enrichment of motifs related to solar-UV signature, the higher frequencies of solar-UV signature related mutations in domesticated accessions than wild accessions, and the enrichment of genes involved in UV damage repair pathway surrounding detected genetic signals of base composition variation together indicate the potential role of solar-UV radiation in driving genome divergence. Motifs related to solar-UV signature having higher frequencies in methylated regions than unmethylated regions establishes the connection between DNA methylation and base-composition variation. The larger [AT]-difference in TE than non-TE regions connects base-composition variation with TEs. The correlations between [AT]-difference and recombination rate indicate recombination may also affect base-composition variation. DNA methylation level, TE density, and recombination rate share similar distribution patterns along the chromosomes, which suggests the likely combined effect of these genomic features on base composition variation. Our findings bring together several components in genome evolution including UV radiation, DNA repair, mutation, DNA methylation, and recombination.

UAV-HTPPs provide great opportunities for large-scale proximal measurements of plant traits. Genotypic differences can be identified through clustering analysis with time series NDVI data obtained from UAV imagery. 1752 diverse maize accessions were classified into two clusters that exhibit different NDVI growth patterns. While NDVI values of accessions in one

cluster either kept increasing across all the surveyed growth stages or reached and stayed at the plateaus at the very last growth stage, NDVI values of accessions in the other cluster increased during the first few growth stages, reached the plateaus, and decreased afterward. Statistical modeling of time series NDVI data to obtain genotype-specific curve parameters that incorporate information from all time points enabled us to study NDVI as a developmental process. GWAS of NDVI curve parameters detected a larger number of loci associated with NDVI than GWAS of NDVI from individual time points, which suggests the advantage of curve parameters over individual time points for the genetic dissection of NDVI. In addition, the dynamic change of SNP effect for the trait associated genetic loci was discovered through GWAS of model fitted NDVI values, which indicate that gene-environment interplay may play an important role in controlling NDVI development. Our analyses demonstrate the great potential of UAV-based remote sensing for genetic dissection of complex traits.

Future Perspectives

Base composition is an essential genome feature. Studies of base-composition evolutionary patterns can advance our understanding of genome evolution. Our study investigated the genome-wide base composition variation pattern in populations separated by a domestication bottleneck event. It will be interesting to expand the current analysis to a large population that consists of exclusively domesticated accessions and contains accessions from different subpopulations separated by different breeding objectives or environments, which may help us reveal how far the base composition difference can extend and what are other components contributing to the base composition variation. The findings in this study make the initial connection between solar-UV radiation and base composition change. GWAS of base composition variation identified a set of genes functions in the UV damage repair pathway. The non-uniform DNA repair was implicated to play a role in generating heterogeneous mutation

patterns and SNP density and divergence between human and chimpanzee (Shendure and Akey, 2015). Sequences encoding DNA repair genes are vulnerable to mutagen attack (Altieri et al., 2008) and the characteristic patterns of different DNA repair genes vary widely (Martincorena and Campbell, 2015). Induced mutation accumulation experiments in model species with contrasting starting materials segregating only at regions surrounding mutation repair genes and UV as the mutagen could be carried out to provide some molecular evidence. Further advances in genomics and epigenomics will also increase our capacity to probe potential connections among base-composition, mutation, DNA repair, and methylation.

Recently UAV-HTPPs have gained more and more interest in crop phenotyping. Our study demonstrated the great potential and effectiveness of the UAV-based remote sensing platform to acquire rapid, detailed NDVI measurements, which in turn facilitate the characterization of NDVI dynamics and the modeling of NDVI growth curves and improve the detection of genetic loci for NDVI in maize. In the future, a large number of growth stages during the growing season should be surveyed with UAV to obtain more comprehensive NDVI measurements, so that more complete NDVI growth curves can be developed to further facilitate the genetic dissection of NDVI. Increasing replications for each genotype is necessary to increase the accuracy of the measurement (Liebisch et al., 2015). Further automation of the image processing and data analysis pipelines with computer vision techniques and machine learning algorithms is also critical, so that more traits such as plant height, canopy cover, green leaf area, vegetation indices, and stress related traits can be easily extracted from UAV imagery and the analysis of phenotypic traits can be efficiently scaled up (Singh et al., 2016; Han et al., 2019; Kitano et al., 2019). Future research effort should also be directed to integrate multi-scale phenotyping data with crop modeling and genomic selection techniques to improve the

prediction accuracy and breeding efficiency (Rutkoski et al., 2016; Crain et al., 2017; Zhao et al., 2019).

References

Altieri F, Grillo C, Maceroni M, Chichiarelli S (2008) DNA damage and repair: from molecular mechanisms to health implications. Antioxid Redox Signal 10: 891-937

Crain J, Reynolds M, Poland J (2017) Utilizing high-throughput phenotypic data for improved phenotypic selection of stress-adaptive traits in wheat. Crop Sci 57: 648-659

Han L, Yang G, Dai H, Xu B, Yang H, Feng H, Li Z, Yang X (2019) Modeling maize above-ground biomass based on machine learning approaches using UAV remote-sensing data. Plant Methods 15: 10

Kitano BT, Mendes CCT, Geus AR, Oliveira HC, Souza JR (2019) Corn plant counting using deep learning and UAV images. IEEE Geosci Remote Sens Lett 58: 1-5

Liebisch F, Kirchgessner N, Schneider D, Walter A, Hund A (2015) Remote, aerial phenotyping of maize traits with a mobile multi-sensor approach. Plant Methods 11: 9

Martincorena I, Campbell PJ (2015) Somatic mutation in cancer and normal cells. Science 349: 1483-1489

Rutkoski J, Poland J, Mondal S, Autrique E, Pérez LG, Crossa J, Reynolds M, Singh R (2016) Canopy temperature and vegetation indices from high-throughput phenotyping improve accuracy of pedigree and genomic selection for grain yield in wheat. G3 6: 2799-2808

Shendure J, Akey JM (2015) The origins, determinants, and consequences of human mutations. Science 349: 1478-1483

Singh A, Ganapathysubramanian B, Singh AK, Sarkar S (2016) Machine learning for high-throughput stress phenotyping in plants. Trends Plant Sci 21: 110–124

Zhao C, Zhang Y, Du J, Guo X, Wen W, Gu S, Wang J, Fan J (2019) Crop phenomics: current status and perspectives. Front Plant Sci 10: 714